BrainLLM-Bench: Model Rankings Transfer Across Listeners but Not Always Across Stimuli or Readouts

Anonymous authors
Paper under double-blind review

Abstract

Reading a brain-likeness leaderboard as a fact about models assumes its ranking survives a change of readout, listeners or text. On one listener's MEG, 32 checkpoints get a reliable leaderboard from representational similarity (ranking reliability 0.98) and an equally reliable one from linear encoding (0.96), in *opposite* orders ($\rho = -0.573$, family-clustered 95% CI $[-0.79, -0.26]$). Brain-likeness benchmarks usually fix one readout and resample only participants, so they can see neither this nor whether a ranking survives a change of text. BrainLLM-Bench changes readout, listeners and stimulus one at a time. Listeners barely move a leaderboard: disjoint listener halves agree at median Spearman ρ of $+0.76$ to $+0.96$ (two crossed cohorts, both readouts). The reversal, for the readouts as conventionally applied, survives the 48 base checkpoints of the registered test under family-clustered uncertainty ($\rho = -0.59$, CI $[-0.78, -0.12]$) but belongs to one cohort, not reproducing on 27-listener MEG or 47-subject English EEG, and feature geometry is not excluded; on the 32 checkpoints it is not determined once static embeddings are also partialled out ($\rho = -0.415$). Text depends on the cohort: on four narratives heard by the same 27 listeners, similarity leaderboards from different stories do not transfer (mean agreement near zero; exploratory — the registered co-primary supports it, the registered dispersion primary did not), while a fully crossed fMRI cohort (25 narratives $\times$ 9 listeners) transfers well ($+0.78$). The benchmark behind these tests is one protocol over five language cohorts (MEG, EEG and intracranial ECoG; English and German; scalp EEG an uncontrolled probe), six leaderboard models and a readout sweep extended from 32 to 57 checkpoints (48 base) spanning 2019–2026. We prove which feature transformations each readout ignores; this shows the two can disagree without either being wrong, not that this is why they disagree here, and the reversal is a disagreement between measurements, not a refutation of brain-likeness. Model size did not consistently order alignment on a Pythia ladder (70M–2.8B) or across three Qwen generations (up to 72B). A brain-likeness ranking is a claim about a readout, a cohort and a stimulus sample, and should name all three.

1 Introduction

Leaderboards that score language models against brain recordings are used to say which model is most brain-like and whether larger ones are more so. Score 32 checkpoints from 12 families against one listener's MEG, over twelve sessions of Sherlock Holmes audiobooks. Representational similarity (Kriegeskorte et al., 2008) returns a leaderboard with ranking reliability 0.98; ridge encoding (Huth et al., 2016) returns an equally reliable one (0.96). The two order the checkpoints in opposite directions ($\rho = -0.573$): similarity puts the autoregressive models last and encoding puts them first.

The readout is usually chosen silently: brain-likeness benchmarks fix one (Schrimpf et al., 2020; 2021), size-scaling results each rest on one (Antonello et al., 2023; Hong et al., 2026), and single alignment numbers are already fragile to positional and word-rate confounds (Hadidi et al., 2026) and to trivial baselines (Feghhi et al., 2024). In vision, which network looks most brain-like already flips with the similarity measure (Soni et al., 2024). Benchmarks put their error bars over participants, but a reader of "model X is more brain-like" assumes the ranking also survives another readout and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

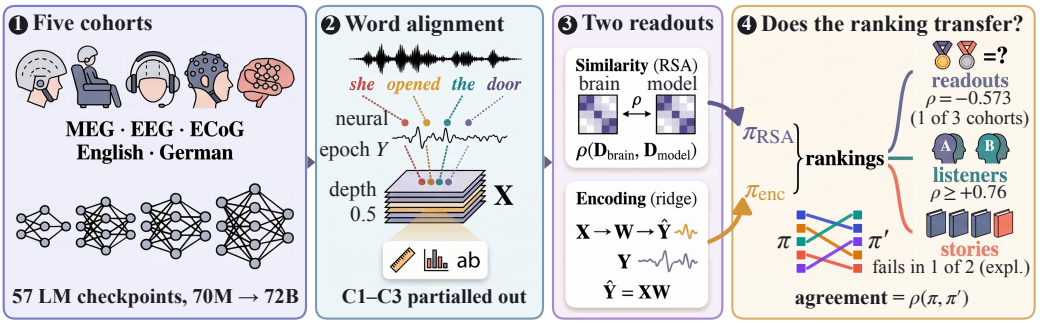


Figure 1: **New listeners leave a ranking in place; a change of readout reversed it in one of three cohorts, and a new story failed to transfer in one of two crossed cohorts (exploratory).** (1) Five language cohorts (three MEG, scalp EEG, intracranial ECoG; English and German) and LM checkpoints from 70M to 72B; the readout comparisons use a 32-checkpoint sweep, extended to 57 scored checkpoints. (2) Each spoken word yields a neural epoch Y and a token state X read at relative depth 0.5. (3) Similarity compares word-by-word dissimilarity matrices, $\rho(D_{\text{brain}}, D_{\text{model}})$, with length, frequency and orthography (C1–C3) partialled out; with each model's static layer-0 embedding (C4) also partialled out, the reversal is not determined ($\rho = -0.415$). Ridge encoding predicts $\hat{Y} = XW$ with train-fold standardization. Each readout, as conventionally applied, ranks the same models (π_{RSA}, π_{enc}). (4) The benchmark asks whether a ranking survives a change of readout, of listener half and of story, scoring agreement as $\rho(\pi, \pi')$; the between-story result is exploratory. Icons, matrices and traces carry no data; the values in (4) repeat Table 1.

another text. Resampling listeners cannot test the second, since every listener heard the same words: the language-as-fixed-effect fallacy (Clark, 1973; Westfall et al., 2016; Yarkoni, 2022), acting on model rankings rather than on effect estimates.

BrainLLM-Bench makes readout, listeners and text explicit axes (Figure 1): it fixes the protocol (statistic, confound battery, ceiling, null) and changes one argument of a ranking at a time (Section 3). Resampling *listeners* (on LibriBrain, sessions) barely moves a leaderboard in all eight cohort, readout and aggregation combinations. The *readout* reversal survives family-clustered inference and the 48 base checkpoints of the registered test. Changing the *text* leaves story leaderboards in place in one crossed cohort and not in the other. What drives the reversal is outside what these data identify: 56% of it is a between-class offset; it does not travel (two other cohorts give positive concordance); feature geometry is not excluded (standardized-feature similarity leaves a family-clustered interval crossing zero), and Proposition 1 shows that readouts can disagree, not why ours do; the claim covers the readouts as conventionally applied, with lexical confounds C1–C3 partialled out, and adding C4 leaves the reversal not determined (Appendix J); and the story failure is exploratory, since its registered primary did not confirm. We contribute:

- **A benchmark that re-scores a ranking with one argument changed.** Five cohorts, six leaderboard models and a 32-to-57-checkpoint readout sweep (2019–2026) under one protocol and both readouts; a new model enters from frozen features, with routes, result files and numerical checks provided for review (Section 4).
- **A transfer test, and what each readout ignores.** Agreement $\rho(\pi, \pi')$ is scored only between determined rankings; we prove which feature transformations each readout ignores, so two readouts need not agree and feature standardization before similarity is the natural geometry control (Proposition 1, Appendix A).
- **Measurements on each axis, and the reporting rule they imply.** Rankings transfer over listeners, but failed to transfer over readouts in one of three cohorts and over stories in one of two crossed cohorts (exploratory) (Table 1), so a ranking should name its readout and stimulus sample; best-cell values overstate alignment (Section 4.3); and passing stored-claim reproduction checks does not establish numerical correctness (Section 5.5).

2 Related Work

Encoding and similarity as readouts of alignment. Encoding models regress each neural response on model features and score held-out prediction (Huth et al., 2016; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Caucheteux et al., 2023; Goldstein et al., 2022); representational similarity analysis compares word-by-word dissimilarity structure without fitting a map (Kriegeskorte et al., 2008; Nili et al., 2014); decoding runs the map in reverse (Tang et al., 2023; Défossez et al., 2023). Diedrichsen & Kriegeskorte (2017) show that encoding, pattern-component and similarity analyses can all be written in terms of the second-moment matrix of the representation and differ in how they evaluate it. We turn that observation into a statement about which feature transformations each readout ignores, and measure what it does to model rankings. Ryskina et al. (2025) added a similarity analysis because encoding barely separated their models; we compare full rankings under both.

Brain-likeness benchmarks. Brain-Score (Schrimpf et al., 2020) and its language counterpart (Schrimpf et al., 2021) score many models against fixed recordings, each neural benchmark under one fixed metric, and Oota et al. (2024) document how far cohorts, statistics and preprocessing vary between studies. In ML evaluation, leaderboard conclusions depend on suite composition (Dehghani et al., 2021) and test-set size (Card et al., 2020), and newer frameworks report coverage and uncertainty explicitly (Liang et al., 2023). These benchmarks treat participants as the sampled unit and fix one readout; we add the readout and the text as axes along which a ranking can move, and report agreement along each.

Scaling of brain alignment. Encoding performance grows roughly log-linearly with model size in fMRI and ECoG studies (Antonello et al., 2023; Gao et al., 2024; Hong et al., 2026), see also Nastase (2025), while other work reports plateaus or saturation (Raugel et al., 2025; Oota et al., 2026). Each rests on one readout; on matched ladders the answer depends on the readout and the ridge range.

Metric dependence, geometry and reliability. In vision, which network looks most brain-like flips with the similarity measure (Soni et al., 2024; Ahlert et al., 2024), and similarity indices differ in the transformations they ignore (Kornblith et al., 2019). Language-model states are anisotropic, with a few "rogue" directions dominating cosine geometry (Ethayarajh, 2019; Timkey & van Schijndel, 2021), and effective dimensionality shapes how well features predict neural data (Elmoznino & Bonner, 2024; Canatar et al., 2023). Noise ceilings bound achievable alignment (Nili et al., 2014); confound analyses show how much of it can be lexical or positional (Hadidi et al., 2026), and under shuffled train–test splits a trivial temporal-autocorrelation feature outscores language models (Feghhi et al., 2024). Our encoding folds hold out a session or listener, except on ds003020 (shuffled TRs within a story and listener, the split they warn about), and we rank models against each other, not against trivial or positional features. We put the confound battery and per-band ceilings inside the score, and test, without a decisive answer, whether feature geometry explains the disagreement.

Stimulus generalization. Treating stimuli as fixed overstates what an experiment generalizes to (Clark, 1973; Westfall et al., 2016; Yarkoni, 2022). Naturalistic encoding studies test on held-out stories (Huth et al., 2016; LeBel et al., 2023), but held-out prediction by one model asks a different question from whether a ranking of models survives a change of story. We measure the latter in two fully crossed cohorts.

3 Problem Setup

Data. A cohort supplies replication units u: sessions of the single LibriBrain listener, listeners elsewhere. In unit u, word i evokes a neural pattern $y_i \in \mathbb{R}^p$, the signal in one cell fixed in advance (region $\times$ band $\times$ window; on LibriBrain, left-temporal sensors, broadband, 300–400 ms), flattened over sensors and time (for encoding, each sensor's mean over the window). Model m reads the same text in sentence context, and $x_i^{(m)} \in \mathbb{R}^{d_m}$ is its state for word i at relative depth 0.5, one layer per model fixed before scoring. Stacking the n_u words gives $Y_u \in \mathbb{R}^{n_u \times p}$ and $X_u^{(m)} \in \mathbb{R}^{n_u \times d_m}$.

Similarity readout. With correlation distance for the brain and cosine distance for the model, $D_{ij}^{\mathrm{b}} = 1 - \mathrm{corr}(y_i, y_j)$ and $D_{ij}^{(m)} = 1 - \langle x_i, x_j \rangle / (\|x_i\| \|x_j\|)$, the unit score is a partial Spearman correlation

over word pairs,

$$s^{\text{RSA}}_{m,u} = \rho_S\big(D^{\text{b}}, D^{(m)} \,\big|\, C_1, C_2, C_3\big), \tag{1}$$

where C_1–C_3 are pairwise word-length and log-frequency differences and orthographic edit distance. With the static layer-0 embedding (C_4) also partialled out, the reversal is not determined (Appendix J).

Encoding readout. Leaving unit u out, features are standardized with training-fold means μ and deviations $\Sigma = \text{diag}(\sigma)$, $\tilde{X} = (X - \mathbf{1}\mu^\top)\Sigma^{-1}$, responses are training-centered, and

$$\hat{W}_\alpha = \arg\min_W \|\tilde{Y}_{-u} - \tilde{X}_{-u}W\|_F^2 + \alpha\|W\|_F^2 = (\tilde{X}_{-u}^\top\tilde{X}_{-u} + \alpha I)^{-1}\tilde{X}_{-u}^\top\tilde{Y}_{-u}, \tag{2}$$

with α chosen by leave-one-out on the training fold from eight log-spaced values from 1e1 to 1e6 (`RidgeCV`; Pedregosa et al., 2011). The unit score is the held-out Pearson correlation averaged over the response coordinates (sensors), $s^{\text{enc}}_{m,u} = \text{mean}_k\,\text{corr}\big(Y_{u,:k}, (\tilde{X}_u\hat{W}_\alpha)_{:k}\big)$.

Rankings and agreement. Readout $r \in \{\text{RSA}, \text{enc}\}$ scores model m on a set of units U by $S^r_m(U) = |U|^{-1}\sum_{u\in U} s^r_{m,u}$, and $\pi^r(U)$ ranks the M models by $S^r(U)$. Two rankings agree to the degree $\rho(\pi,\pi') = 1 - 6\sum_m(\pi_m - \pi'_m)^2/\big(M(M^2-1)\big)$. The paper changes one argument at a time:

$$\underbrace{\rho\big(\pi^{\text{RSA}}(U), \pi^{\text{enc}}(U)\big)}_{\text{readout}}, \qquad \underbrace{\rho\big(\pi^r(U_1), \pi^r(U_2)\big)}_{\text{listener halves } U_1 \sqcup U_2 = U}, \qquad \underbrace{\rho\big(\pi^r_k, \pi^r_l\big)}_{\text{stories } k \neq l}. \tag{3}$$

Here π^r_k is computed on story k alone with all listeners. Listener halves resample people with the words fixed; story pairs resample words with the people fixed, and only the latter asks whether a ranking generalizes to new text. Two rankings are compared only if each is *determined*, i.e. reproduces across disjoint halves of its own units (the rankability precondition). Checkpoints are not independent draws: the 32-checkpoint sweep falls into 12 families. The LibriBrain readout comparisons carry a checkpoint-level Fisher-z interval and a family-clustered bootstrap interval (10,000 draws over whole families), and claims about them are held to the clustered one; for MEG-MASC and Brennan only the Fisher-z interval is stored, and it is labelled as such.

Proposition 1 (What each readout ignores)**.** *Fix the neural data and confounds.* ***(i)*** *Replacing X by cXQ with $c > 0$ and Q orthogonal leaves $D^{(m)}$, hence s^{RSA}, unchanged; a translation $x_i \mapsto x_i + b$ or an unequal rescaling of coordinates changes $D^{(m)}$ in general.* ***(ii)*** *Replacing X by $XA + \mathbf{1}b^\top$ with A positive diagonal leaves $\tilde{X}$, hence s^{enc} at every α, unchanged. If $\tilde{X}_{-u}$ has full column rank, then as $\alpha \to 0$ the held-out predictions are unchanged for every invertible A and every b.* ***(iii)*** *If $\tilde{X}_{-u} = \sum_k s_k u_k v_k^\top$ is an SVD, the training-fold fitted values are $\sum_k \frac{s_k^2}{s_k^2+\alpha}\,u_k u_k^\top \tilde{Y}_{-u}$.* ***(iv)*** *If $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues of the centered feature covariance and $\text{PR} = (\sum_k \lambda_k)^2/\sum_k \lambda_k^2$ its participation ratio, the leading component carries at least a fraction $1/\text{PR}$ of the squared Frobenius norm of the centered Gram matrix $\bar{X}\bar{X}^\top$.*

Proofs are in Appendix A. Read together: similarity sees the feature geometry with each direction weighted by its variance, so in a model with participation ratio near one the word-pair geometry is essentially that of one direction (iv); encoding ignores per-coordinate scale and offset (ii) and admits each direction with a weight that saturates at one once $s_k^2 \gg \alpha$, whatever its variance share (iii). The two readouts are functions of different summaries of the same spectrum, the split Kornblith et al. (2019) draw between CKA-type and CCA-type similarity indices, so their rankings can disagree without either being wrong. This fixes three design choices: we report both readouts rather than merging them; feature standardization before similarity (which (ii) leaves invisible to encoding) is the natural geometry control; and the rank-one perturbation diagnostic must refit encoding, because a sample-varying added direction changes the column space that (ii) protects. The proposition says the readouts *can* order models differently; it does not say that this is why they do on our data (Section 5.3). Part (iv) concerns the centered inner-product geometry; the cosine distance of Eq. (1) is taken on uncentered features and divides out each word's norm.

4 Benchmark

4.1 Cohorts and Models

Five cohorts enter (Appendix Table 11). LibriBrain (Özdogan et al., 2025) is one listener; we use its 36 Sherlock 1 MEG recordings (12 sessions × three story parts); MEG-MASC (Gwilliams et al.,

2023) is English MEG (10 subjects in the tensor archive, 27 in the raw rebuild for crossed and readout analyses); MEG-SCANS (Habersetzer et al., 2025) is German MEG from 23 of its 24 listeners (one lacks usable word-aligned epochs); Brennan (Brennan & Hale, 2019) is scalp EEG whose public release (Brennan, 2018) has 49 listeners, kept as an uncontrolled probe (47 in rebuilt analyses: two subjects' public files are corrupt); and the "Podcast" ECoG dataset (Zada et al., 2025) has nine patients, all in natural comprehension. Table 1's fMRI cohort ds003020 (LeBel et al., 2023) is an extra crossed cohort, not one of the five; we use OpenNeuro release 3.1.1 (LeBel et al., 2026), whose nine listeners are one more than the eight the data paper describes. Six models form the per-cohort leaderboard (Appendix Table 4): GPT-2 and GPT-2-medium (Radford et al., 2019) and BERT-base (Devlin et al., 2019) on the English MEG and ECoG cohorts, and German-GPT-2 (Bayerische Staatsbibliothek (dbmdz), 2020), German-BERT (Chan et al., 2019) and multilingual-BERT (Devlin et al., 2019) on the German one. The readout comparison ranks a 32-checkpoint sweep over 12 families and three objectives (autoregressive, masked, encoder–decoder; Appendix Table 3). Registered extensions reach 57: 48 base checkpoints, which alone enter the registered reversal test, and 9 post-trained ones (exploratory only). 18 of the 25 added checkpoints were first scored on text without inter-word spaces; all 25 were re-extracted and the results below use those features (Appendix D).

4.2 Alignment Tensor and Fixed Protocol

The tensor route computes Eq. (1) in every $(\text{region}, \text{band}, \text{window}, \text{layer})$ cell, one tensor per cohort, Fisher-z averaged over units for reporting (constants in Appendix D). It partials out all four confounds, including the model's static layer-0 embedding (C4), under which the fixed-depth reversal of Section 5.3 is not determined. A per-band split-half ceiling, in the spirit of the RSA noise ceiling (Nili et al., 2014), bounds how much reproducible word-level structure each band carries. Inference treats the unit as the replication unit in a mixed-effects model over a 27-cell grid fixed in advance, checked against a word-label permutation null; the grid excludes broadband, whose result rests on per-unit reliability instead (Appendix K). Each confirmatory question has one registered primary with its decision rule fixed before scoring, so we apply no family-wise correction across questions; the grid is the only larger test family (Appendix D).

4.3 Leaderboard Optimism and Harness Use

Each value in a per-cohort leaderboard (Appendix Table 4) is the best cell of a search, read from the same data that chose it. Choosing the cell on held-out units instead costs 85% of the value across nine ECoG patients and 6% across the 36 recordings of one MEG listener; subsampling one cohort shows unit count drives much of that gap, with heterogeneity contributing a residual (Appendix E). Best-cell values are optimistically selected in-sample estimates, most optimistic in small cohorts of dissimilar units, so the benchmark defines its primary metric at a pre-registered cell; ceiling-normalized scores are undefined where the ceiling is not reliably positive. A new model enters either route from frozen features (`bench.run`, `p4_concordance32.py`); ridge fits a readout and never fine-tunes the model. A comparison must name route, cell, controls and resampling unit, and the tensor route's validation does not validate the fixed-depth one (Appendix Table 2).

5 Results

We change the arguments of Eq. (3) one at a time: listeners, text, readout (Table 1).

5.1 Stability Across Listeners

Whether a leaderboard is determined depends on the uncertainty of the *difference* between two models, not the marginal uncertainty of either. Resampling participants moves every model together: for the encoding readout 77–91% of each model's marginal standard error is this shared component, and it cancels in every pairwise comparison. Judged by the marginal error, the leaderboards look unidentifiable; judged correctly, the between-model spread exceeds the paired-difference error severalfold and rankings survive participant resampling in all eight LibriBrain and Brennan readout $\times$ aggregation combinations (Table 1 caption).

Table 1: **Rankings survive new listeners, but not always a new story or a new readout: agreement $\rho(\pi, \pi')$ when one argument of Eq. (3) changes.** *Top:* medians over listener-half splits (ds003020: median of per-story medians) or story pairs; the gap and partition rows give a difference with its 95% interval. MEG-MASC story means are -0.025 (full) and $+0.003$ (equal length); no ds003020 story pair of 300 is negative; unit halves on LibriBrain and Brennan give Kendall $\tau = 0.89$–0.97, with between-model spread 2.4–7.4$\times$ the paired-difference SE. *Bottom:* readout concordance with 95% intervals, [ckpt] Fisher-z over checkpoints, [fam] family-clustered bootstrap over 12 families. †Exploratory. ‡Registered test that held (32 checkpoints: concordance below the reliability ceiling; 48 base: family-clustered interval below zero); the C4 row is a registered test that left the reversal not determined (Appendix J). §Shuffled five-fold TR splits within each story and listener (Section 2), which can raise encoding agreement. ¶Checkpoints first scored on unspaced text re-extracted (Appendix D); the top panel and the per-story row still include three such checkpoints. At equal length one of the four story leaderboards is not determined (F3).

	MEG-MASC (4 stories $\times$ 27)		ds003020 fMRI (25 $\times$ 9)	
What changes	RSA	encoding	RSA	encoding§
Listener halves	$+0.896$	$+0.764$	$+0.831$	$+0.961$
Stories, full length	$-0.411^{\dagger}$	$+0.674^{\dagger}$	$+0.778$	$+0.937$
Stories, equal length	$-0.146^{\dagger}$	$+0.635^{\dagger}$	–	–
Listener − story gap	–	–	$+0.053$	$+0.024$
95% interval, stories resampled			$[+0.012, +0.109]$	$[+0.006, +0.034]$
Same − other story (partitions)	$+0.564^{\dagger}$	$+0.034^{\dagger}$	–	–
95% interval	$[+0.35, +0.78]$	$[-0.03, +0.09]$		

Readout concordance on	$\rho(\pi^{\text{RSA}}, \pi^{\text{enc}})$	95% interval or note
LibriBrain, 32 checkpoints‡	-0.573	$[-0.77, -0.28]$ [ckpt]; $[-0.79, -0.26]$ [fam]
LibriBrain, 32, C4 also partialled out	-0.415	$[-0.67, -0.08]$ [ckpt]; $[-0.73, +0.01]$ [fam]
LibriBrain, 48 base checkpoints‡¶	-0.59	$[-0.78, -0.12]$ [fam]
LibriBrain, class means removed	-0.25	$p = 0.16$; 56% of $\lvert\rho\rvert$ is between classes
MEG-MASC, 27 listeners¶	$+0.446$	$[+0.12, +0.69]$ [ckpt]
Brennan EEG, 47 listeners¶	$+0.366$	$[+0.02, +0.63]$ [ckpt]
MEG-MASC, each story alone	-0.147 to $+0.384$	sign not determined

5.2 Transfer Across Stories

MEG-MASC presents four narratives and all 27 listeners heard all four, a crossed design. Holding cell, preprocessing, checkpoints and listeners fixed, we score all 32 models within each story; each story's trials span both recording sessions of every listener, so story is not confounded with session. Every story's similarity leaderboard is determined on its own, yet, exploratorily, the leaderboards do not agree with each other (Figure 2a): the four stories form two pairs that agree internally, and all four cross-pair agreements are negative. The median of these six values (Table 1) overstates a bimodal set; the mean sits near zero. The data therefore support a ranking that does not transfer to another story, not one that inverts. Capping every story at the shortest length (registered control F4) removes most of the disagreement (Table 1), and at that length one story (The Black Willow) no longer meets the rankability precondition (reliabilities in Appendix M); by the registered reading, the full-length disagreement is at least partly a power effect. Encoding is more stable than similarity but stays below the listener axis (Figure 2b).

Pre-registration outcome. This test was registered before scoring (ADR 0005) and was *not confirmed*. The registered primary, a story-to-listener dispersion ratio with threshold > 2, came out below one for both readouts (Appendix M), which also fired falsifier F1 (ratio ≤ 1), whose pre-committed reading we state as written: on that statistic, narratives move the leaderboard *no more than listeners do*. The registered co-primary, pairwise between-story agreement, met its criterion. We therefore report the transfer failure as *exploratory*; why the two statistics disagree, and what a confirmatory test needs, is in Appendix M.

A second crossed cohort bounds the effect. On ds003020, where nine listeners heard 25 narratives, every story's leaderboard is determined and the listener-minus-story gap replicates in *direction*

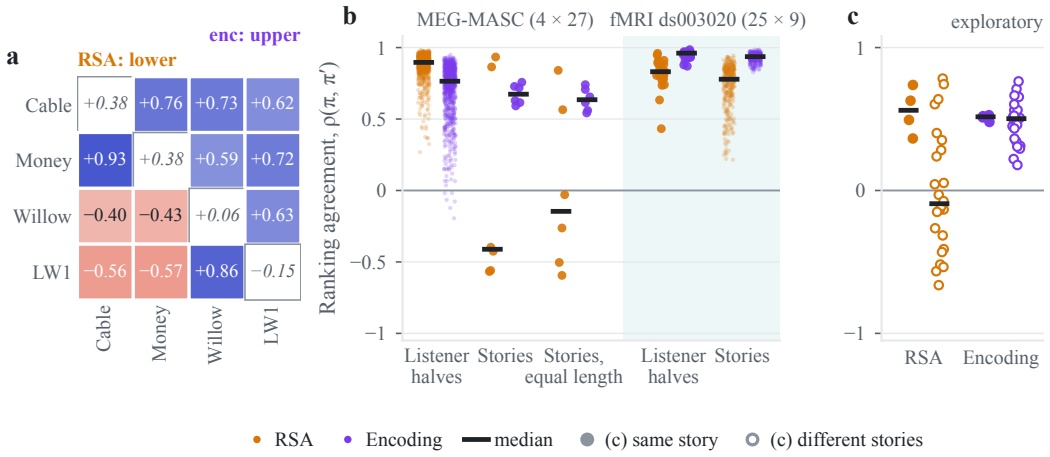


Figure 2: **New listeners leave the leaderboard in place in both crossed cohorts; a new story does so on fMRI but not, under similarity, on MEG-MASC (exploratory).** (a) MEG-MASC story-by-story ranking agreement (exploratory): similarity below the diagonal, encoding above; the diagonal gives each story's within-story readout concordance. Under similarity the stories form two internally agreeing pairs. (b) Agreement for each change of argument, both readouts, in the two crossed cohorts: listener halves (MEG-MASC, 4 stories × 200 splits; ds003020, per-story medians), story pairs (MEG-MASC 6 pairs at full and equal length; ds003020 300 pairs). Bars are medians (the text quotes means for (c)); repeated splits and overlapping pairs are dependent, not independent observations. (c) Same-story versus different-story equal-sized partitions (exploratory: one of eight partitions is rankable). All marks are recomputed from the stored `crossed_narrative` and `ds003020` receipts, which still include three unspaced checkpoints (Appendix D).

for both readouts (encoding on shuffled within-story TR folds, which can inflate agreement), with story-resampled intervals above zero. Its magnitude does not: between-story agreement is high and no story pair disagrees in sign (Table 1), so rankings largely transfer there. What the two cohorts jointly support is conditional: a listener-stable leaderboard can fail to transfer across stimuli, it did, exploratorily, in one of two crossed cohorts, and only measuring it tells which case a benchmark is in. Four stories are a small sample of narratives, and part of MEG-MASC's spread is story-sampling noise, as the registered ratio suggested.

What the design cannot separate. Words and narratives vary together here, so we write *stimulus-sample dependence* and never claim a narrative-specific leaderboard. A matched-size partition split separates the two only suggestively: similarity leaderboards from the same story agree more than those from different stories, encoding shows no gap, and only one of eight partitions meets the rankability precondition (Figure 2c, Appendix M). Two crossed cohorts show that transfer is heterogeneous but not which dataset properties predict it, and encoding needs far more data than similarity before a ranking exists: on eight Brennan segments of about 8 000 words, similarity is determined on all and encoding on one (Appendix M).

5.3 Dependence on the Readout

On LibriBrain the two readouts rank the 32 checkpoints in *opposite* orders ($\rho = -0.573$). Both rankings are individually reliable (0.98 and 0.96 against an attenuation ceiling of 0.97), so this is a disagreement between determined rankings. The family-clustered interval, over 12 units rather than 32, is wider and still excludes zero; with the added checkpoints re-extracted, the reversal also survives the 48-checkpoint base set that spans 2019–2026 (Table 1), so the registered falsifier that it was an artifact of older models did not fire.

The disagreement is about architecture *classes* more than individual models (Figure 3a). Removing each class's mean from both readouts leaves a weak within-class concordance, so 56% of the effect is a between-class offset: similarity puts autoregressive models last and encoding puts them first, a gap

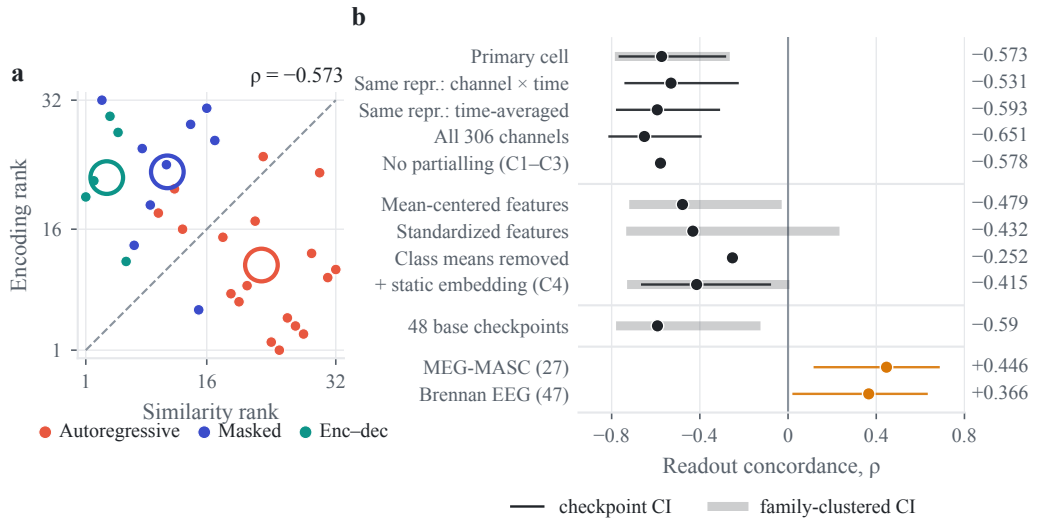


Figure 3: **The reversal is mostly between classes and survives representation and cell changes; its sign is unresolved after full feature standardization or C4.** (a) Similarity rank against encoding rank for the 32 LibriBrain checkpoints (rank 1 strongest); large rings mark each class's mean rank, and the dashed diagonal is equal ranking. (b) Readout concordance under each variant, LibriBrain in black and the other two cohorts in amber. Thin bars: Fisher-z intervals over checkpoints; thick bars: family-clustered bootstrap intervals (12 families) where computed; a point without a bar has no stored interval; hairlines separate groups of variants. Sources: `concordance32_CORRECTED` and its variants; `concordance32_lb_allch_CORRECTED` (both readouts on all channels; the ledger's mixed-cell entry is not plotted); `correction_2026_08_derived` (no partialling); `concordance32_c4` (C4); and, re-extracted after a tokenization defect (Appendix D), `adr0006_registered_verdicts_spacefix` and the MEG-MASC and Brennan receipts.

of $+25.7$ ranks that survives permuting the architecture label over whole families. Autoregressive checkpoints are also wider, but with width matched the gap shrinks and still survives the family-clustered test (Appendix O). Objective class also fixes context direction and token pooling and, on LibriBrain, the tokenization rule, which changes 15 of the 18 autoregressive feature sets little; no test here separates these from class (Appendix D). Neither representation nor cell explains it, and dropping the lexical partialling does not either: handing both readouts the identical neural representation, scoring all 306 channels or dropping C1–C3 leaves the concordance close to its headline value (Figure 3b). With the static-embedding control C4 added it is -0.415, with a family-clustered interval that includes zero: neither the reversal nor its change is determined (Table 1; Appendix J).

Feature geometry is not excluded. Proposition 1 predicts where to look: cosine similarity weighs the dominant direction of anisotropic autoregressive models by its variance, while standardized ridge does not see per-coordinate scale. Standardizing features before similarity keeps the point estimate negative, but under the headline's own inference the mean-centered interval barely excludes zero and the fully standardized one crosses it (Figure 3b). We therefore do not claim the reversal is independent of feature geometry; the claim is that the readouts *as conventionally applied* disagree. A post hoc refit of ten models confirms parts (i)–(ii) on real data (centering changes similarity ranks and leaves encoding scores unchanged to numerical precision) and shows that a sample-varying rank-one addition does move encoding ranks. The old claim that standardization absorbs that perturbation was false, and the synthetic perturbation does not identify the endogenous cause (Appendix O). Effective dimensionality covaries with the readout gap within the autoregressive class only, which is exploratory. A third readout, contrastive decoding, gives no determined ranking (Appendix G).

The reversal belongs to one cohort. Under the identical protocol (same band, window, depth and controls; right temporal sensors on MEG-MASC, all 60 channels on Brennan), MEG-MASC (27 listeners, both readouts determined) and Brennan EEG (47 listeners) give *positive* concordance (on Brennan the interval only just excludes zero), and the class gap is near zero on Brennan (Appendix O;

before re-extraction). Recording depth does not explain the difference (11 of 12 LibriBrain sessions reverse on their own, Appendix I), neural measurement quality does not either, and two further pre-registered accounts were refuted (Appendix F). This is an existence proof, not a field-wide constant.

5.4 Additional Findings

Scaling, retested at modern scale (registered before extraction). Across three Qwen generations (Qwen et al., 2024; Yang et al., 2025; Qwen Team, 2026), 2.5 (0.5B–72B), 3 and 3.5, in no generation does size order both readouts positively on the primary cell (Appendix Figure 4). On the largest ladder the ordering is itself readout-dependent: encoding falls with size (72B vs 0.5B, paired $p = 0.0005$) while similarity rises ($p = 0.001$). A post hoc, training-fold-expanded ridge grid attenuates the encoding size correlation from -0.964 to -0.464, so that slope is regularization-sensitive, while 72B stays below 0.5B (Appendix Table 7). On MEG-MASC's 27 listeners the same rungs do not order similarity by size, and the encoding arm fails its rankability gate, so it is not determined. The six-point Pythia ladder (70M–2.8B), scored at the fixed depth, does not order encoding by size either (Appendix L). Over the whole range, size-scaling claims made under one readout (Antonello et al., 2023; Hong et al., 2026) do not describe both; across six base–instruct twins instruction tuning shifts neither readout consistently ($2/6$, $4/6$; exploratory). **A band comparison** (Appendix K): broadband reaches a higher peak alignment than 30–80 Hz gamma in 206 of 213 recordings; on LibriBrain, where the ceiling was measured, gamma has no reproducible word-level structure to align to. The 70–150 Hz intracranial envelope is not tested. **Ten cross-setting hypothesis tests** (Appendix L) are conceptual extensions and analogies, not a replication-rate estimate; one row lacks its source tensors and is unverified.

5.5 Fragility of Score Computation

A separate class of instability concerns how scores are computed: errors of this class produce plausible numbers rather than exceptions. Three occurred in this pipeline (Appendix N) with one cause, code written as if inputs were on a unit scale and applied to neural data in physical units. The least visible, a 10^{-12} guard in a correlation denominator, shrank stronger-predicting models less and changed the readout concordance of an earlier analysis in this project from -0.573 to -0.678; changing only that constant recovers -0.573. Our claim-reproduction harness read $97/97$ green throughout, because it checks stored results against the paper's claims and cannot see that they came from defective code. Reproducibility and correctness are different properties, and the first is routinely offered as evidence of the second (Eklund et al., 2016; Botvinik-Nezer et al., 2020).

6 Discussion and Limitations

Across listeners rankings were stable; across readouts they reversed, as conventionally applied, on the single LibriBrain listener (feature geometry not excluded) but not on two other cohorts; across stories they held in one crossed cohort and, exploratorily, failed in the other. A claim that a model is more brain-like should rest on both readouts, with standardized features as a geometry check, on determined rankings with family-clustered intervals, and on a named cohort and stimulus sample, with uncertainty over stimuli where the design allows; otherwise it is a claim about one readout, cohort and text. None of this certifies a mechanism (Murphy, 2026): the band result is a reliability-grounded preprocessing correction, every scorecard verdict is our operationalization, and the stimulus result shows rankings can fail to generalize across stimulus samples, not that the narrative causes it.

Single-subject where it is load-bearing. LibriBrain, the primary cohort, is one listener over many sessions, and it is where the broadband ceiling is computed. MEG-MASC reproduces the ceiling ordering, but its word-level dissimilarity sits near the noise floor, leaving the normalization's alignment side ill-defined: clean on one cohort, echoed on a second, not on all five. *Magnitude scale and in-sample selection.* Raw best-cell alignment is 10^{-3}–10^{-2} in MEG and $\approx 2 \times 10^{-2}$ in ECoG, and headline values select cells in-sample at a 6–85% cost (Section 4.3) with no held-out rebuild. *Scope of the readout comparison.* With C4 added, the fixed-depth reversal is not determined (CI including zero). The cross-cohort contrast also crosses the tokenization rule of autoregressive features and the encoding fold (Appendices D, I); unspaced-input checkpoints were re-extracted except in ds003020 and other MEG-MASC and Brennan analyses.

AI Use Statement

We used generative AI tools (Anthropic Claude and OpenAI Codex agents) to implement experiment code and check it for correctness, to collect public datasets, to check references, and to polish the writing. The authors take full responsibility for all content.

Ethics Statement

BrainLLM-Bench is built entirely on previously published human-recording datasets, each collected under its own institutional ethics approval and participant consent: LibriBrain, MEG-MASC, MEG-SCANS, the Brennan EEG corpus, and the "Podcast" intracranial dataset (OpenNeuro `ds005574`), whose original data papers we cite; the crossed-narrative analysis also uses the public fMRI dataset ds003020 (LeBel et al., 2023). We did not collect new human data. We will redistribute only derived artifacts (per-word word lists, sensor-region maps and alignment tensors), never raw recordings, and the word lists come from public narrative stimuli (audiobooks, read stories and podcasts) that carry no participant-identifying content. The intracranial cohort is the most sensitive; we use it only through the maintainers' de-identified OpenNeuro release and report electrode locations only at the MNI-region level. Brain–LLM alignment benchmarks carry a modest dual-use consideration, that better alignment metrics could inform neural decoding of language; the effects here are small and word-level, well short of message reconstruction, and we flag the concern rather than dismiss it.

Reproducibility Statement

The quantitative headline results trace to files under the project's `results/` directory; the unavailable baseline-comparison tensors in the historical scorecard are an explicit exception. We verify stored summaries rather than rerun all analyses: the per-cohort headline cells and the broadband count to `cross_cohort_summary.tsv` and `per_unit_bb_vs_gamma.tsv` (four cohorts are scored under the full confound-controlled tensor protocol, and Appendix Figure 7 reports exactly those four; Brennan EEG has no headline cell under the tensor protocol, so it is absent from that figure, and it is scored on the fixed-depth route, Table 1 and §5.3), the per-band reliability ceiling to `p2_perband_ceiling.json`, the cross-setting outcomes to `scorecard/scorecard_transport.tsv` (legacy labels in `scorecard/scorecard.tsv`), the family-clustered intervals of §5.3 to `p4_encoding/family_clustered_ci_CORRECTED.txt` and, for the 48 base checkpoints, `p4_encoding/space_fix_20260925/aggregate/adr0006_registered_verdicts_spacefix.json` (the unspaced run: `concordance_modern_final_v2.json`), and the family map to `codes/model_families.py`, the single definition every clustered inference imports. The protocol is deterministic given a fixed random seed for the permutation nulls. The tensor harness takes public recordings plus word lists and sensor-region maps. Its validation recomputes all 504 stored per-(model, session) values behind the LibriBrain cell to within 10^{-9} from cached neural dissimilarity matrices, so it checks the model side, the confound partialling and the statistic but not the neural preprocessing, and it has a portable known-answer test. This validates the tensor route, not the C1–C3 fixed-depth sweep. That sweep's scores are reproducible from the cached features, and 31 of its 32 checkpoints also by re-extraction: the 17 autoregressive ones with the earlier extractor revision shipped in the supplementary material, the 14 others with the documented one (Appendix J); GPT-2 small is matched by file identity. Re-extracting the autoregressive ones with the documented revision does not reproduce them, because it tokenizes differently (Appendix D). Of the 25 extension checkpoints, 18 were extracted with the documented extractor under a library version at which, for their tokenizers, it removes the spaces between words (Appendix D); all 25 were re-extracted, with the receipts and gates in `p4_encoding/space_fix_20260925.json`. An anonymized subset for review is provided in the supplementary material: the harness and its stored-result checks (193 registered claims; the run marks the 33 that pin superseded values kept for provenance, 25 of them values computed on the unspaced features, and the 48 that belong to a separate boundary-effect analysis not part of this submission), the checker for Table 1, the derived result files and scorecard cited above, the thirteen pre-registration documents (ADR 0001–0010, one of them an amendment; the readout-concordance test; the C4 control; the re-extraction), the sensor-region atlases and the analysis scripts. The alignment tensors, word lists, cached model features and cached neural dissimilarity matrices are not in it, so the 504-value validation cannot be rerun from it; they will be released upon publication, with the public repository and an archival DOI, which are withheld here only to preserve double-blind anonymity.

References

Jannis Ahlert, Thomas Klein, Felix Wichmann, and Robert Geirhos. How aligned are different alignment metrics? *arXiv preprint arXiv:2407.07530*, 2024.

Richard Antonello, Aditya Vaidya, and Alexander G. Huth. Scaling laws for language encoding models in fMRI. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019. doi: 10.1038/s41586-019-1119-1.

Bayerische Staatsbibliothek (dbmdz). German GPT-2 model (dbmdz/german-gpt2). `https://huggingface.co/dbmdz/german-gpt2`, 2020.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Ilina Bhaya-Grossman, Matthew K Leonard, Yizhen Zhang, Laura Gwilliams, Keith Johnson, Junfeng Lu, and Edward F Chang. Shared and language-specific phonological processing in the human temporal lobe. *Nature*, 649(8095):140–151, 2025. doi: 10.1038/s41586-025-09748-8.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. In *ICML*, 2023.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow. Zenodo, version 1.0, doi: 10.5281/zenodo.5297715, 2021.

Rotem Botvinik-Nezer et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582:84–88, 2020.

Jonathan R. Brennan. EEG datasets for naturalistic listening to "Alice in Wonderland". University of Michigan Deep Blue Data, doi: 10.7302/Z29C6VNH, 2018.

Jonathan R. Brennan and John T. Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, 14(1):e0207741, 2019.

Abdulkadir Canatar, Jenelle Feather, Albert Wakhloo, and SueYeon Chung. A spectral theory of neural prediction and alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, 2020.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441, 2023. doi: 10.1038/s41562-022-01516-2.

Branden Chan, Timo Möller, Malte Pietsch, and Tanay Soni. German BERT (bert-base-german-cased). `https://huggingface.co/google-bert/bert-base-german-cased`, 2019.

Xupeng Chen, Ran Wang, Amirhossein Khalilian-Gourtani, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nature Machine Intelligence*, 6(4):467–480, 2024.

Herbert H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359, 1973.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*, 2020.

Nathan E. Crone, Alon Sinai, and Anna Korzeniewska. High-frequency gamma oscillations and human brain mapping with electrocorticography. In *Event-Related Dynamics of Brain Oscillations*, volume 159 of *Progress in Brain Research*, pp. 275–295. Elsevier, 2006. doi: 10.1016/S0079-6123(06)59019-3.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5: 1097–1107, 2023.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508, 2017. doi: 10.1371/journal.pcbi.1005508.

Anders Eklund, Thomas E. Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016.

Eric Elmoznino and Michael F. Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20(1):e1011792, 2024. doi: 10.1371/journal.pcbi.1011792.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 55–65, 2019.

Ebrahim Feghhi, Nima Hadidi, Bryan Song, Idan Asher Blank, and Jonathan C. Kao. What are large language models mapping to in the brain? A case against over-reliance on brain scores. *arXiv preprint arXiv:2406.01538*, 2024.

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. Scaling, but not instruction tuning, increases large language models' alignment with language processing in the human brain. *bioRxiv preprint 2024.08.15.608196*, 2024.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations (ICLR)*, 2019.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, et al. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, 16(1):10529, 2025. doi: 10.1038/s41467-025-65518-0.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 10(1):862, 2023. doi: 10.1038/s41597-023-02752-5.

Till Habersetzer, Svea Steuer, Andreas Radeloff, and Bernd T. Meyer. MEG-SCANS – a comprehensive magnetoencephalography speech dataset with stories, chirps and noisy sentences. *Scientific Data*, 12(1):1938, 2025. doi: 10.1038/s41597-025-06397-4.

Nima Hadidi, Ebrahim Feghhi, Bryan H. Song, Idan A. Blank, and Jonathan C. Kao. Spurious alignment between large language models and brains can emerge from non-robust methods and overlooked confounds. *Nature Communications*, 17:5769, 2026. doi: 10.1038/s41467-026-72253-7. Earlier preprint bioRxiv 10.1101/2025.03.09.642245, "Illusions of alignment between large language models and brains emerge from fragile methods and overlooked confounds".

Linyang He, Tianjun Zhong, Richard Antonello, Gavin Mischler, Micah Goldblum, and Nima Mesgarani. Far from the shallow: Brain-predictive reasoning embedding through residual disentanglement. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL `https://arxiv.org/abs/2510.22860`.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhuoqiao Hong, Haocheng Wang, Zaid Zada, Harshvardhan Gazula, David Turner, Bobbi Aubrey, Leonard Niekerken, Werner Doyle, Sasha Devore, Patricia Dugan, et al. Larger language models better align with neural representations of natural language. *eLife*, 13:RP101204, 2026. doi: 10.7554/eLife.101204.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4110–4124, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*, 2020.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10:555, 2023.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. An fMRI dataset during a passive natural language listening task. OpenNeuro, dataset ds003020, version 3.1.1, doi: 10.18112/openneuro.ds003020.v3.1.1, 2026.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020. doi: 10.18653/v1/2020.acl-main.703.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Meta. Llama-3.2-1B. Hugging Face model card, `https://huggingface.co/meta-llama/Llama-3.2-1B`, 2024.

Omer Moussa and Mariya Toneva. Brain-tuning improves generalizability and efficiency of brain alignment in speech models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.21520.

Elliot Murphy. No country for old linguists: LLM-brain alignment underdetermines neural computation. *arXiv preprint arXiv:2609.03160*, 2026.

Samuel A. Nastase. Larger language models better align with the reading brain. *Nature Computational Science*, 5:994–995, 2025. doi: 10.1038/s43588-025-00905-7.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553, 2014. doi: 10.1371/journal.pcbi.1003553.

Subba Reddy Oota, Zijiao Chen, Manish Gupta, Raju S. Bapi, Gael Jobard, Frederic Alexandre, and Xavier Hinaut. Deep neural networks and brain alignment: Brain encoding and decoding (survey). *Transactions on Machine Learning Research*, 2024.

Subba Reddy Oota, Vijay Rowtula, Satya Sai Srinath Namburi, Khushbu Pahwa, Anant Khandelwal, Manish Gupta, Tanmoy Chakraborty, and Bapi S. Raju. Linguistic properties and model scale in brain encoding: from small to compressed language models. *arXiv preprint arXiv:2602.07547*, 2026.

Miran Özdogan, Gilad Landau, Gereon Elvers, Dulhan Jayalath, Pratik Somaiya, Francesco Mantegna, Mark Woolrich, and Oiwi Parker Jones. LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2025.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. doi: 10.3115/v1/D14-1162.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Joséphine Raugel, Stéphane d'Ascoli, Jérémy Rapin, Valentin Wyart, and Jean-Rémi King. Scaling and context steer LLMs along the same computational path as the human brain. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2512.01591.

Supratim Ray and John H. R. Maunsell. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology*, 9(4):e1000610, 2011. doi: 10.1371/journal.pbio.1000610.

Maria Ryskina, Greta Tuckute, Alexander Fung, Ashley Malkin, and Evelina Fedorenko. Language models align with brain regions that represent concepts across modalities. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2508.11536.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020. doi: 10.1016/j.neuron.2020.07.040.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Ansh Soni, Sudhanshu Srivastava, Marvin Maechler, Konrad Kording, and Meenakshi Khosla. Conclusions drawn from neural network to brain alignment depend strongly on the chosen similarity measure. *bioRxiv*, 2024. doi: 10.1101/2024.08.07.607035.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4527–4546, 2021.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8:544–561, 2024.

Jacob Westfall, Thomas E. Nichols, and Tal Yarkoni. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*, 1:23, 2016.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tal Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1, 2022. doi: 10.1017/S0140525X20001685.

Zaid Zada, Samuel A Nastase, Bobbi Aubrey, Itamar Jalon, Sebastian Michelmann, Haocheng Wang, Liat Hasenfratz, Werner Doyle, Daniel Friedman, Patricia Dugan, et al. The "Podcast" ECoG dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12 (1):1135, 2025. doi: 10.1038/s41597-025-05462-2.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Jiajie Zou, David Poeppel, and Nai Ding. Constituent-constrained word prediction during language comprehension. *Nature Neuroscience*, 29(6):1498–1509, 2026. doi: 10.1038/s41593-026-02272-6.

A Proof of Proposition 1

Throughout, the neural patterns, the brain dissimilarity D^{b} and the confound matrices C_1–C_3 do not depend on the model features, so a readout score is unchanged whenever the model-side object it is computed from is unchanged.

(i) For $c > 0$ and orthogonal Q, $\langle cQ^\top x_i, cQ^\top x_j\rangle = c^2\langle x_i, x_j\rangle$ and $\|cQ^\top x_i\| = c\|x_i\|$, so every cosine, hence $D^{(m)}$ and the partial Spearman correlation of Eq. (1), is unchanged. For the converse it suffices to exhibit one case. For $x_1 = (1,0)$, $x_2 = (0,1)$, $x_3 = (2,1)$ the cosine distances are $d_{12} = 1$, $d_{13} = 1 - 2/\sqrt{5} \approx 0.106$, $d_{23} = 1 - 1/\sqrt{5} \approx 0.553$, so $d_{13} < d_{23} < d_{12}$. Translating by $b = (0,2)$ gives $d_{12} \approx 0.106$, $d_{13} \approx 0.008$, $d_{23} \approx 0.168$, so $d_{13} < d_{12} < d_{23}$; rescaling by $\mathrm{diag}(1,3)$ gives $d_{23} \approx 0.168 < d_{13} \approx 0.445 < d_{12} = 1$. Both maps change the rank order of the word-pair dissimilarities, so, embedded in a larger word set with a suitable D^{b} and confounds, they change the partial Spearman statistic.

(ii) Let $X' = XA + \mathbf{1}b^\top$ with $A = \mathrm{diag}(a)$, $a_j > 0$. On the training fold the column means and deviations become $\mu' = A\mu + b$ and $\sigma' = A\sigma$, so $\tilde{X}' = (XA + \mathbf{1}b^\top - \mathbf{1}\mu^\top A - \mathbf{1}b^\top)A^{-1}\Sigma^{-1} = \tilde{X}$ on both training and held-out rows. The standardized design is identical, so the selected α, $\hat{W}_\alpha$ and the held-out predictions are identical at every α. (The implementation adds 10^{-8} to σ; the identity then holds up to that guard.) Now let A be any invertible matrix and write $Z = \tilde{X}_{-u}$, $Z_u = \tilde{X}_u$. Centering commutes with right multiplication, and per-column scaling is right multiplication by a positive diagonal matrix, so $Z' = ZB$ and $Z'_u = Z_u B$ with $B = \Sigma A \Sigma'^{-1}$ invertible. If Z has full column rank, the $\alpha \to 0$ limit of Eq. (2) is the least-squares solution and

$$Z'_u(Z'^\top Z')^{-1}Z'^\top \tilde{Y}_{-u} = Z_u B\, B^{-1}(Z^\top Z)^{-1}B^{-\top}B^\top Z^\top \tilde{Y}_{-u} = Z_u(Z^\top Z)^{-1}Z^\top \tilde{Y}_{-u},$$

so the held-out predictions, and the Pearson score computed from them, are unchanged.

(iii) With $Z = \sum_k s_k u_k v_k^\top$, $Z^\top Z + \alpha I = \sum_k (s_k^2 + \alpha) v_k v_k^\top + \alpha P_\perp$, where $P_\perp$ projects onto the orthogonal complement of the right singular vectors; since $Z^\top \tilde{Y}_{-u}$ lies in their span, the training-fold fitted values are $Z\hat{W}_\alpha = \sum_k s_k u_k v_k^\top \sum_l (s_l^2 + \alpha)^{-1} s_l v_l u_l^\top \tilde{Y}_{-u} = \sum_k \frac{s_k^2}{s_k^2 + \alpha} u_k u_k^\top \tilde{Y}_{-u}$.

(iv) Let $\bar{X}$ be the centered features, n words. The nonzero eigenvalues of the Gram matrix $K = \bar{X}\bar{X}^\top$ are those of $\bar{X}^\top\bar{X} = n\,\mathrm{Cov}$, namely $n\lambda_k$, and $K = \sum_k n\lambda_k v_k v_k^\top$ with orthonormal v_k, so the rank-one terms are orthogonal in the Frobenius inner product and $\|K\|_F^2 = n^2 \sum_k \lambda_k^2$. The leading term's share is $\lambda_1^2 / \sum_k \lambda_k^2$. Because $0 \le \lambda_k \le \lambda_1$, $\sum_k \lambda_k^2 \le \lambda_1 \sum_k \lambda_k$, which gives both $\lambda_1^2 / \sum_k \lambda_k^2 \ge \lambda_1 / \sum_k \lambda_k$ and $1/\mathrm{PR} = \sum_k \lambda_k^2 / (\sum_k \lambda_k)^2 \le \lambda_1 / \sum_k \lambda_k$. Chaining the two, the share is at least $1/\mathrm{PR}$. □

Scope. Parts (i)–(iii) are exact statements about the pipeline of Section 3, except that (ii)'s second claim is an $\alpha \to 0$ limit, whereas the scored α is selected from a grid. Part (iv) concerns the centered inner-product geometry; cosine distance is taken on uncentered features and divides by each word's norm, so (iv) motivates, and does not prove, the reading that a low-participation-ratio model's similarity score is dominated by one direction. None of the four parts says which transformation separates the models in our data; that question is the empirical one of Appendix O, and it remains open.

B Benchmark Context and Protocol Matrix

B.1 Prior Work and Benchmark Comparison

Brain–LLM alignment studies. A large body of work relates language-model features to brain responses through encoding models (Schrimpf et al., 2021; Caucheteux et al., 2023; Antonello et al., 2023; Toneva & Wehbe, 2019; Goldstein et al., 2022), representational similarity (Kriegeskorte et al., 2008; Nili et al., 2014), and decoding (Tang et al., 2023; Défossez et al., 2023). In vision, conclusions flip with the similarity measure (Soni et al., 2024; Ahlert et al., 2024). These efforts differ in cohort, modality, model, statistic and preprocessing, so their magnitudes sit on incompatible scales (Oota et al., 2024).

Analysis	Scoring and layer rule	Controls and validation scope
Original tensor	Partial RSA across layer/cell grid; raw argmax shown descriptively	C1–C4; per-band ceilings and 27-cell inference grid. The portable tensor validation checks this route.
Fixed-depth sweep	RSA and held-out ridge encoding; relative depth 0.5	RSA: C1–C3 (C1–C4 for the 32, Appendix J); encoding: train-fold standardization. 32 then 57 checkpoints; model-family clustered comparisons.
Crossed stimuli	Same fixed-depth scores, separately by stimulus (ds003020: word features averaged per TR)	C1–C3 RSA (ds003020: C1–C2); encoding leaves a listener out (ds003020: shuffled five-fold TR splits within a story and listener); listener and stimulus resampling reported separately. MEG transfer result exploratory.

Table 2: Executed analysis routes. C1–C3 are length, frequency and orthography; C4 is model-specific static embedding. With C4 added, the 32-checkpoint fixed-depth reversal is not determined (Appendix J); the 57-checkpoint sweep and the MEG-MASC crossed-stimulus scores use C1–C3; ds003020 uses C1–C2.

Existing benchmarks. The integrative-modeling line (Schrimpf et al., 2021) scores many models against fixed brain recordings and is the closest prior benchmark in spirit, but it is single-metric (encoding), single-modality within each dataset, and carries no built-in confound battery or per-band reliability ceiling; brain-tuning and driving-response studies (Moussa & Toneva, 2025; Tuckute et al., 2024) likewise fix one readout. Because the language benchmarks above each fix a single readout, none reports whether readouts agree on a ranking of language models (in vision, Soni et al. (2024) and Ahlert et al. (2024) show that they need not) — the question our readout comparison asks; the ML-side leaderboard-reliability literature that motivates treating participants and stimuli as two sampled axes is reviewed in Appendix P. BrainLLM-Bench differs on four axes: it scores similarity and encoding side by side on its primary cohort (and reports why decoding does not yet yield a determined ranking), so their concordance is measured rather than assumed; it spans modalities (MEG, EEG, ECoG) and languages within one protocol; it carries confound control and reliability ceilings inside the score rather than as add-ons; and it treats the frequency band as a benchmarked dimension rather than a fixed preprocessing choice.

The reliability critique. The confound and reproducibility critiques (Hadidi et al., 2026; Feghhi et al., 2024) directly motivate a standardized frame. We take their lesson, that positional, word-rate and other simple stimulus features must be controlled before an alignment score is read, and build confound control and clustered inference into the measurement rather than adding them post hoc: every tensor cell is scored with the confound battery available and every null is clustered; raw best-cell values are reported separately and labelled as such.

B.2 Executed Route Matrix

Table 2 lists the executed analysis routes, what each controls, and which one the portable validation covers.

C Model-Size Ladder Visualization

Figure 4 shows the three Qwen ladders of Section 5.4, one row per readout.

Summary of Findings

The appendices below report in full the findings summarized in the main text. Figure 5 collects the six-family readout diagnostic behind Appendix G and the corrected Pythia ladder of Appendix L.

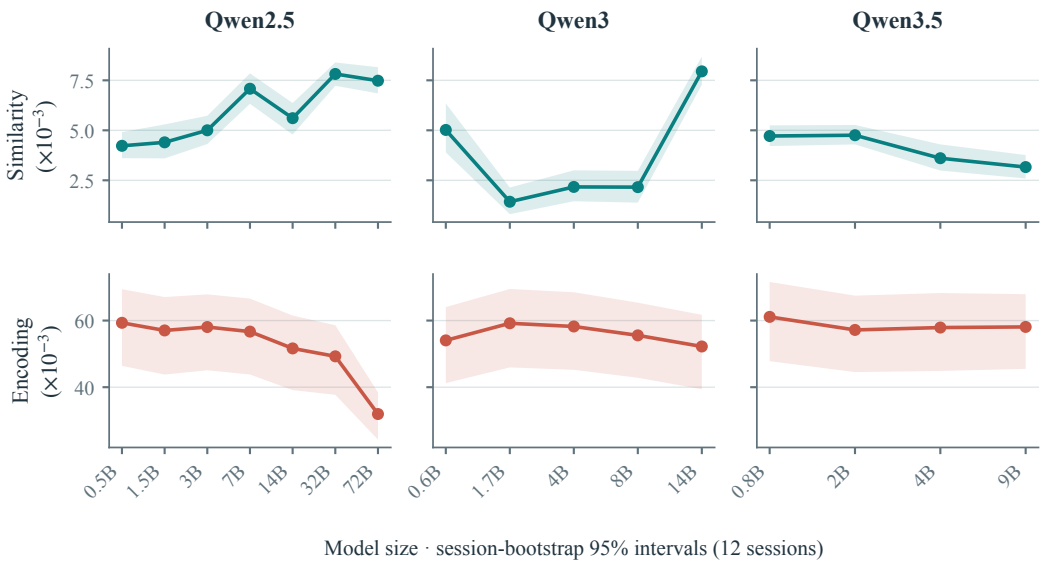


Figure 4: **Three model-size ladders, separated by readout.** Similarity (top) and encoding (bottom) use shared scales within each row; the two rows have different ranges. Points are per-session means, ribbons are session-clustered bootstrap 95% intervals (12 sessions). Each generation retains its published rungs. Values are read from concordance_spacefix_libribrain.json (every extension rung re-extracted after the tokenization defect, Appendix D); the registered tests, rather than visual slope alone, determine the claims. These curves use the original bounded 1e1–1e6 ridge grid; expanded-grid sensitivity is reported separately in the text.

D Protocol Parameters

Sessions are subsampled to at most 2500 words under a fixed seed before the dissimilarity matrix is built, which bounds the dissimilarity vector at 12.5 MB and is what keeps the tensor tractable. The per-band reliability ceiling is computed on a separate, smaller repeated-word cache (462 words over 12 sessions), since a split-half ceiling needs words that recur across sessions. Word epochs run from -200 to $+800$ ms around onset. Each epoch is filtered into six bands with a fourth-order Butterworth filter: broadband 1–80, delta 1–4, theta 4–8, alpha 8–13, beta 13–30, and gamma 30–80 Hz. Sensors are partitioned into six anatomical wedges (frontal, temporal, parieto-occipital $\times$ left, right) by fixed sensor-layout atlases for the Neuromag-306 and KIT-208 systems (in the bundle); scalp EEG uses all channels; ECoG uses language regions of interest (ROIs) in MNI (Montreal Neurological Institute) standard space.

Test family and multiplicity. Each confirmatory question in the main text — the concordance-versus-ceiling test, the crossed-narrative dispersion test, the scaling ladders, and the modern-set reversal — has one registered primary statistic with a decision rule fixed before scoring, so its confirmatory test family has size one and we apply no family-wise correction across questions; no error-rate guarantee is therefore claimed across questions, and an exploratory analysis does not rescue a registered test that failed. Within a question, the only test family larger than one is the 27-cell LME grid, controlled at Benjamini–Hochberg FDR (Benjamini & Hochberg, 1995) $q = 0.05$ over the 27 cells within each cohort $\times$ model. Every other p-value in the paper is descriptive, and an analysis labelled *exploratory* (the story-transfer finding whose registered primary did not confirm, the participation-ratio mechanism, and the base–instruct twins) carries no confirmatory weight, whatever its p-value. The family map is a single module imported by every clustered inference (`codes/model_families.py`); it folds DistilBERT and ALBERT into BERT, since they share its corpus and tokenizer lineage, giving 12 families rather than the 14 a name-level split would give — fewer resampling units and wider clustered intervals, which is the conservative side when the claims rest on those intervals excluding zero. The registration documents are internal and shipped in the supplementary material. Only the

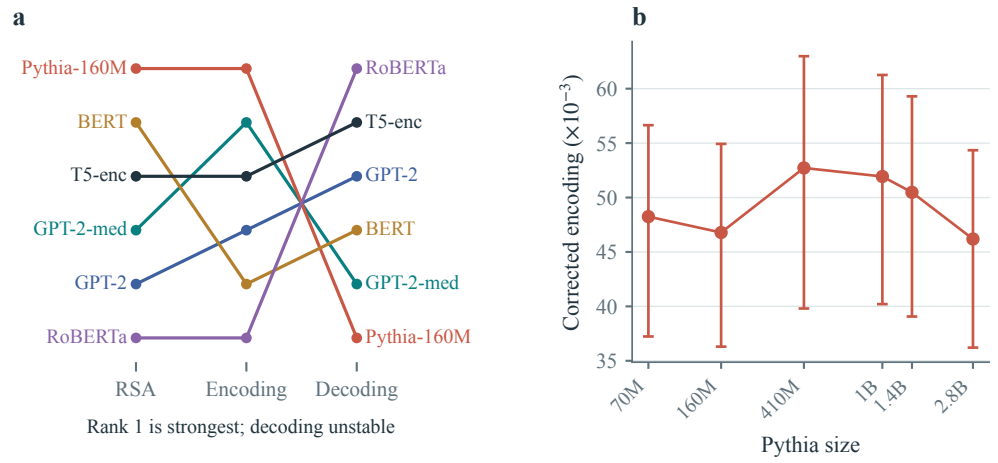


Figure 5: **Readout-dependent ranks and a receipt-backed size ladder.** (a) The six-family legacy diagnostic, with decoding ranked by the mean of twelve leave-one-session-out retrains. The decoding order is not stable (split-half reliability 0.54); RSA and encoding reliability are both 0.94. This legacy diagnostic predates the correction of Appendix N and was not recomputed with it. (b) The corrected six-rung Pythia encoding ladder at preregistered depth, with session-bootstrap 95% intervals. This replaces the old illustrative three-point peak-normalized panel, whose RSA values were hand-entered. Source: concordance.json, decoding_lb_persess.json, scaling_ladder_test_CORRECTED.json.

Table 3: The 32 checkpoints of the readout comparison, by objective class and family; the 12 families are the resampling units of every family-clustered interval (`codes/model_families.py`). All are the public pretrained releases.

Class	Family	Checkpoints
Autoregressive (18)	GPT-2 (Radford et al., 2019)	small, medium, large, XL
	Pythia (Biderman et al., 2023)	70M, 160M, 410M, 1B, 1.4B, 2.8B
	OPT (Zhang et al., 2022)	125M, 350M, 1.3B
	GPT-Neo (Black et al., 2021)	125M, 1.3B
	Llama (Grattafiori et al., 2024)	Llama-3.2-1B (Meta, 2024)
	Qwen (Qwen et al., 2024)	Qwen2.5-0.5B, Qwen2.5-1.5B
Masked (9)	BERT (Devlin et al., 2019)	base, large; DistilBERT (Sanh et al., 2019); ALBERT-base-v2 (Lan et al., 2020)
	RoBERTa (Liu et al., 2019)	base, large
	DeBERTa (He et al., 2023)	v3-base, v3-large
	ELECTRA (Clark et al., 2020)	base discriminator
Encoder–decoder (5)	T5 (Raffel et al., 2020)	small, base, large
	BART (Lewis et al., 2020)	base, large

32-checkpoint registration predates the reversal (it was written when a six-family pilot existed); the later ones, including the 48-checkpoint, C4 and re-extraction registrations, were written after it was known, so those that concern the reversal test the robustness of a known effect in the same listener and do not replicate it. Table 3 lists the 32 checkpoints by class and family.

Compute. Everything downstream of feature extraction runs on one CPU workstation; no analysis stage needs a GPU. Language-model feature extraction for the 32-checkpoint sweep is the only GPU stage, embarrassingly parallel over (model, narrative) pairs; features are extracted once and cached, and every result in this paper is computed from those caches.

Feature-extraction provenance. Every checkpoint is read at the same relative depth, but the reading procedure depends on the class (`extract_llm_layers_multi.py`): causal models read the

prefix that ends in the target word and are read at its last token; masked models read the whole sentence, including the words after the target, and the target's sub-tokens are averaged; encoder–decoder models are read the same way through the encoder stack alone. Objective class therefore also fixes context direction and token pooling, so the between-class component of the reversal (Section 5.3) cannot be attributed to the training objective alone. Among causal models, what differs is how the prefix is tokenized. Two rules were used. The *joined-string* rule passes the prefix as one string with the tokenizer's default settings; it wrote the LibriBrain features of the 18 autoregressive checkpoints of the 32-checkpoint sweep (17 by an earlier revision of the documented extractor `extract_llm_layers_multi.py`, shipped in the supplementary material, and GPT-2 small by the older single-model `extract_llm_layers.py`). The *word-list* rule of the documented revision (`is_split_into_words`, `add_prefix_space`) was introduced to fix target-word alignment in the masked and encoder–decoder models; it wrote the LibriBrain features of those 14 checkpoints, of all 25 checkpoints of the registered extensions (16 base, 9 post-trained), and of every checkpoint on MEG-MASC, Brennan and ds003020 (the 27- and 47-listener analyses re-index those features rather than re-extracting them). This attribution rests on the extraction code and job history, not on recomputed features. For the GPT-2 tokenizer, and Pythia's under a nearby library version, the two rules differ only in whether the sentence-initial word carries a leading space. For the Qwen and Llama tokenizers under transformers 4.57.1, the library version of most extraction runs, the word-list rule inserts no space between words: in two LibriBrain test batches every non-initial word lost its leading space ($252/252$ word boundaries for each of the 18 Qwen and Llama tokenizers probed; the other checkpoints extracted under that version share those tokenizer files), so the model reads, for example, "Intheyeareighteen…"; the read-out token is still the last sub-token of the target word, in its space-less form. The Qwen3.5 checkpoints and one other post-trained checkpoint were extracted under transformers 5.15.0, where the same probe loses no space ($0/252$), so they read spaced text. The 14 masked and encoder–decoder tokenizers lose no space under this rule (0 of 46,511 word boundaries each, in 2,000 LibriBrain sentences). Rerunning the documented extractor on the 30,000 scored LibriBrain words leaves the 14 masked and encoder–decoder rows unchanged (median row cosine 1.0000000 for all 14) and changes the autoregressive rows: per checkpoint, 36% (GPT-Neo-125M) and 93–99.8% (the other 17) of rows fall below cosine 0.9999 with the cached rows, and the median row cosine is 0.9965–0.9999 for the 15 GPT-2, Pythia, OPT and GPT-Neo checkpoints and 0.786 (Llama-3.2-1B), 0.855 and 0.897 (Qwen2.5-0.5B, -1.5B), whose cached rows come from spaced text and whose re-extracted rows from unspaced text. Three consequences follow. (i) On LibriBrain the rule is collinear with objective class among the 32 checkpoints, so the between-class component of the reversal (Section 5.3) is not separated from it. (ii) The autoregressive checkpoints were tokenized under the joined-string rule on LibriBrain and under the word-list rule elsewhere, and Llama-3.2-1B and Qwen2.5-0.5B/1.5B first read unspaced text on MEG-MASC, Brennan and ds003020 (including the listener and story analyses of Table 1) but not on LibriBrain, so the cross-cohort contrast also crosses the two rules. (iii) Of the 25 extension checkpoints, 18 were first scored on unspaced text: 12 of the 16 base checkpoints added in the 48-checkpoint test (Qwen2.5 3B–72B, Llama-3.1-8B and -70B, the five Qwen3 base checkpoints) and 6 of the 9 post-trained ones, so both members of five of the six base–instruct twins; the Qwen3.5 ladder and its twin were not. The Qwen3 ladder, all seven rungs of the Qwen2.5 ladder on MEG-MASC and the 3B–72B rungs of the LibriBrain Qwen2.5 ladder were first scored on unspaced text, while the LibriBrain 0.5B and 1.5B rungs were not, so on those features the 72B-versus-0.5B contrast compared the two input conditions. *Re-extraction.* Under a pre-registration written after the defect was found and before any re-extracted feature existed (`PREREG_space_fix_reextract_20260925.md`), all 25 extension checkpoints (18 to repair the defect, 7 to put every autoregressive LibriBrain checkpoint under one rule), the MEG-MASC Qwen2.5 ladder and the three small checkpoints on MEG-MASC and Brennan (36 extractions) were re-extracted with the joined-string rule. All 36 passed the tokenization gate (every row decodes to the spaced text) and the scorer-identity gate (the scorer reproduces every published per-session value on the cached features); the 25 LibriBrain extension extractions, whose weight load also produced the MEG-MASC ladder rows, passed a weight-identity gate (the documented rule on the same loaded weights reproduces the cached rows), and for the three small checkpoints the new code path reproduced their cached LibriBrain features bit for bit. Every registered outcome below uses the re-extracted features. Each follows the pre-registered rule (`p4_encoding/space_fix_20260925.json`). On the unspaced features, for comparison: the 48-checkpoint test gave $\rho = -0.58$ (family-clustered CI $[-0.78, -0.14]$; re-extracted -0.59, $[-0.78, -0.12]$); the LibriBrain Qwen2.5 ladder gave encoding $\rho = -0.89$ and similarity $+0.61$ (-0.96 and $+0.93$); the Qwen3 ladder's registered label was "mixed" (now "not sup-

ported"; re-extracted, the per-generation test gives, with session-clustered 95% CIs, Qwen2.5 similarity $\rho = +0.93$ $[+0.82, +0.96]$ and encoding -0.96 $[-1.00, -0.82]$, Qwen3 $+0.30$ $[+0.30, +0.40]$ and -0.40 $[-0.70, 0.00]$, Qwen3.5 -0.80 $[-1.00, -0.80]$ and -0.20 $[-1.00, -0.20]$, and no generation is "supported"; with this few rungs the session bootstrap of a rank correlation takes few values, so an interval endpoint can equal the point estimate); the base–instruct twins gave $3/6$ and $4/6$ $(2/6,\ 4/6)$; the MEG-MASC and Brennan concordances were $+0.489$ $[+0.17, +0.72]$ and $+0.342$ $[-0.01, +0.62]$ ($+0.446$ $[+0.12, +0.69]$ and $+0.366$ $[+0.02, +0.63]$; the Brennan interval now excludes zero, narrowly). A descriptive check in the pre-registration expected the 7 extension checkpoints that already read spaced text to change by a median row cosine of at least 0.99; 6 of the 7 changed by more (median row cosine 0.978–0.986; the unspaced ones 0.851–0.919), more than the first-word difference alone produced for the GPT-2-type checkpoints, so by the pre-registered rule their classification as spaced is not confirmed by this check and rests on the tokenizer probe alone; all 7 were re-extracted, so no reported number depends on it. ds003020 and the MEG-MASC story, listener-half and crossed-narrative analyses were not re-run, nor were the repeat-removal analysis (Appendix F), the Brennan segment reliabilities (Appendix M) and the Brennan class gap (Appendix O); all still include the three unspaced checkpoints; the other 15 autoregressive checkpoints on MEG-MASC, Brennan and ds003020 keep the word-list rule. C4 was not rerun. The 32-checkpoint LibriBrain headline read spaced text throughout and is unchanged. Re-extracting the 17 autoregressive checkpoints with the shipped earlier revision reproduces their cached rows (Appendix J, Deviation 2); re-extracting them with the documented revision would not reproduce the reported fixed-depth scores exactly.

The 27-cell inference grid is three left-hemisphere regions (temporal, frontal, parieto-occipital) $\times$ three bands (delta, theta, gamma) $\times$ three layer groups (phonological, syntactic, semantic). It is a deliberate subset of the full tensor, fixed in advance.

E Full Per-Cohort Leaderboard

Table 4 is the per-cohort leaderboard, a descriptive summary rather than the paper's headline finding (that is the readout-dependence of §5.3): the best-aligning cell for each cohort $\times$ model, with its region, band, layer, latency, and the count of mixed-effects survivors. Two facts stand out. First, absolute alignment is small and modality-ordered: MEG best-ρ ranges 0.002 to 0.013, and intracranial ECoG is the largest at 0.015 to 0.022, the expected invasive advantage (Appendix Figure 7). Second, the winning band is broadband in 11 of the 12 cohort $\times$ model cells; the only cell that peaks elsewhere is LibriBrain-BERT at delta, a neighboring low-frequency band. The peak latency spans 275 to 768 ms post-onset, with 9 of the 12 cells between 340 and 590 ms, and the peak region is lateral temporal in every MEG cohort, so the benchmark reproduces the lateral-temporal, mid-latency word-alignment cell that the field expects, now under one protocol across the four cohorts carried through it. A per-model leaderboard on the LibriBrain headline cell (Appendix Figure 8) orders the three architectures scored on LibriBrain (best band per model) by raw alignment; BERT-base leads at 13.1×10^{-3}, above the three-times-larger GPT-2-medium (6.9×10^{-3}), a size-versus-architecture anomaly that Appendix L returns to. This ordering is on raw alignment and is descriptive: the confound-controlled partial that gates significance is near-null at the group mean (§K), so the leaderboard ranks models by an uncontrolled quantity and should be read as "which model aligns most before confound control," not as a validated brain-likeness score. The distinction is not cosmetic: BERT tops this raw leaderboard, but under the confound-controlled similarity readout of the 32-model protocol the highest-ranked models are encoder–decoder and masked ones (T5-small at 11.1×10^{-3}, ahead of DeBERTa and BART), with the autoregressive family last — the class ordering of §5.3, not this one. So the leaderboard's winner is confound-control-dependent — a separate axis from the assay-dependence of §5.3, pointing the same way: the ranking is a property of the measurement choice as much as of the model.

A third dependence is larger than either, and we measure it rather than assert it. The best cell in Table 4 is chosen on the same data its value is read from, so the printed ρ is an optimistically selected quantity. We price that on the two cohorts whose tensors we can rebuild, choosing the cell on every unit but one and reading it on the held-out unit.

The cost depends on the cohort, not the statistic. Across nine ECoG patients with different electrode coverage, held-out alignment collapses from 9.6 to 1.4×10^{-3} — about 85% is selection, and BERT-

Cohort	Model	best ρ	region	band / layer	latency (ms)	LME sig (pos/neg)
LibriBrain	GPT-2	0.0061	temporal_L	broadband / L5	346	13/0
LibriBrain	GPT-2-medium	0.0069	temporal_L	broadband / L7	346	13/0
LibriBrain	BERT-base	0.0131	temporal_L	delta / L0	447	18/0
MEG-MASC	GPT-2	0.0040	temporal_R	broadband / L11	768	0/6
MEG-MASC	GPT-2-medium	0.0039	temporal_R	broadband / L17	768	6/8
MEG-MASC	BERT-base	0.0025	temporal_R	broadband / L4	467	10/0
MEG-SCANS	German GPT-2	0.0043	temporal_R	broadband / L12	567	6/0
MEG-SCANS	German BERT	0.0053	temporal_R	broadband / L0	527	4/0
MEG-SCANS	mBERT	0.0055	temporal_R	broadband / L4	587	8/1
ECoG	GPT-2	0.0216	PCG	broadband / L10	451	n/a
ECoG	GPT-2-medium	0.0203	PCG	broadband / L16	340	n/a
ECoG	BERT-base	0.0154	PCG	broadband / L5	275	n/a

Table 4: Per-cohort headline cells. Each cell is selected on the same data its value is read from, so these are optimistically selected in-sample estimates, not held-out ones (§4.3). For each cohort $\times$ model, the best-aligning tensor cell with its region, band, layer, latency, and count of FDR-controlled linear mixed-effects (LME) survivors out of the 27 region $\times$ band $\times$ layer-group inference cells of §4.2 (ECoG has no session-level LME). Region names follow Appendix D; PCG is precentral gyrus. Broadband is the winning band in 11 of 12 cells; ECoG carries the largest raw alignment. Source: `results/cross_cohort_summary.tsv`.

base's held-out estimate is slightly negative, indistinguishable from no alignment. Across thirty-six recordings of one MEG listener it costs almost nothing, 6.4 to 6.0×10^{-3}, about 6%. Which cell wins explains the gap: on LibriBrain, in the rebuilt 864-cell grid used for this held-out test, all three models pick left-temporal broadband, beside the pre-registered cell, and leave-one-out reselects it in nearly every fold, whereas across patients the argmax moves with the electrodes. Restricting the LibriBrain search to k random cells shows this is not a search-space effect — the penalty stays near zero from $k = 16$ to $k = 864$. The cohorts differ in grid and word budget, so we compare the in-sample-to-held-out ratio within each, not absolute values across them.

Unit *count* is a competing explanation the search-space control does not address, since it varies the cells searched rather than the units averaged. Varying count directly, by subsampling MEG-SCANS by subject with everything else fixed, the tax rises steeply as units are removed: 13.3% at twenty-three subjects, 34.4% at nine, 53.2% at four (Spearman $\rho = -0.96$ against subject count, $p < 0.001$). Count therefore reproduces much of the 85%-versus-6% contrast. A residual survives — at a matched nine units ECoG still pays 85% against 34.4% — so heterogeneity contributes too, but our cohorts vary both at once and cannot separate them. The supported claim is the weaker one: best-cell values are optimistically selected estimates, most optimistic in small cohorts of dissimilar units, which describes the small intracranial studies where this reporting style is most common. It is why the primary metric is defined at a pre-registered cell rather than the argmax.

Two honest reads of Table 4 belong with it. The survivor counts split by direction, and for MEG-MASC GPT-2 all six FDR survivors are negative-direction, so its positive best-cell is not backed by positive population evidence; two further cells are mixed rather than clean. And the best-ρ printed is the raw Spearman, whereas the confound-controlled partial before ceiling normalization is near zero at the LibriBrain group mean (-0.0036: broadband, C1–C4 partialled out, GPT-2 layer 5 on the 462-word ceiling cache of Table 6; a different quantity from the fixed-depth scores of the readout comparison). That most published-looking cells do not survive the stringent yardstick is itself a benchmark finding.

F Two Pre-Registered Accounts of the Reversal

Two accounts of why the reversal appears on LibriBrain and not elsewhere were pre-registered before the data existed, with their falsifiers fixed in advance, and both were refuted. We report them as refuted rather than dropping them, because a reader has no way to tell a hypothesis that was never tested from one that was tested and failed.

Concordance as a function of neural reliability. The registered prediction was that MEG-MASC would fall strictly inside $(-0.68, +0.34)$ and be more negative than Brennan. It landed at $+0.489$,

outside and on the wrong side ($+0.446$ with its three unspaced checkpoints re-extracted, still outside). The accompanying falsifier also fired: on a common estimator the three cohorts' word-level RDM reliabilities are 0.774 (LibriBrain), 0.786 (MEG-MASC) and 0.858 (Brennan) — a range of 0.08 against concordances spanning 1.06 (1.02 with the three checkpoints re-extracted), and not monotone in reliability.

The reversal induced by removing stimulus repeats. LibriBrain presents each word once, while MEG-MASC measures each stimulus token in 27 subjects and Brennan in 47. The registered test averaged k subject-measurements per token and walked k down to 1, reproducing LibriBrain's regime by construction. The reversal was not induced: at $k = 1$ the cross-validated-distance concordance is $+0.140$ (Brennan) and $+0.191$ (MEG-MASC), positive in both. The premise was weaker than we thought in any case, for a reason visible only afterwards — the per-group RDM is computed *within* a subject, so cross-subject repeats never enter it, which is why LibriBrain and MEG-MASC show the same RDM reliability despite 1 against 27 measurements per token.

The one constructive result from that arm survives its refutation: cross-validated neural distances raise leaderboard reproducibility from 0.664 to 0.947, but spend half the data doing it. Matched on the number of subjects actually used, the plain estimator wins at two subjects (0.371 against 0.072) and the cross-validated one wins from four upward ($0.678/0.544$, $0.832/0.690$, $0.947/0.664$). The numbers in this section come from the analysis run log; their per-dose result files are not in the review bundle.

G Contrastive-Decoding Readout

A learned decoder is trained stochastically, and that makes its model ranking unstable in a way the two representational readouts are not (Appendix Figure 5). Its split-half ranking reliability is 0.54, 90% interval $[0.09, 0.83]$, against 0.94 for both representational readouts on the same six families (Figure 5); retraining on the identical held-out session with different seeds gives a mean pairwise rank reliability of 0.17 over six seeds (`results/p4_encoding/decoding_seedvar.json`). Its apparent relation to the other readouts moves from $+0.06$ to -0.77 depending only on whether it is scored from a single best-layer run or from the mean of twelve leave-one-session-out retrains, and one lucky run puts GPT-2 first at $13.1\times$ chance while across twelve retrains GPT-2 averages $5.6\times$ and never ranks first.

Under the precondition stated in §5.1 this readout does not produce a determined ranking, so a comparison between it and either representational readout would be a correlation involving an ordering that is not itself identified. We report it here and exclude it from the concordance analysis.

H Separability of the Ranked Models

A reordering between two rankings is only informative if the systems being ranked are distinguishable in the first place. §5.1 answers this for the 32-checkpoint panel using paired-difference resampling; the family-level tests below were computed earlier and point the same way.

Under representational similarity the six families are separably different (Friedman $\chi^2 = 41.2$, $p = 9 \times 10^{-8}$ over 12 sessions; 11 of 15 pairwise differences significant at uncorrected $p < 0.05$), and the ranking they induce is reproducible across sessions (split-half Spearman median 0.94, 90% interval $[0.89, 1.00]$). Linear encoding, scored with per-session leave-one-session-out held-out R, gives a ranking that is just as stable: split-half Spearman median 0.94, 90% interval $[0.83, 1.00]$ over the 924 balanced session splits (`results/p4_encoding/families_lb_persess.json`, `concordance_separability.json`).

The alignment magnitudes are small in absolute terms. The per-family best-cell, positive-direction values sit between 0.003 and 0.016, above the fixed-cell group-mean partial of -0.0036 reported for broadband in §K (GPT-2 layer 5, 462-word ceiling cache), which averages over the same negative cells the leaderboard discussion flags; the two figures are not in tension.

I Depth and the Readout Reversal

The reversal is measured on LibriBrain, 36 recordings from 12 sessions of one listener, and does not appear on cohorts contributing one session per subject. Depth per unit and cohort identity are therefore confounded in the comparison as reported, and two readings survive it: that shallow per-subject recordings cannot resolve a disagreement deep ones can, or that the reversal is specific to this dataset. They differ in what they predict when depth is varied *within* LibriBrain, which holds cohort, protocol, models and cells fixed and changes only the data behind each estimate.

Re-estimating every model from k randomly chosen sessions instead of twelve (400 draws per k, using the stored per-session alignment matrices, so this re-averages the paper's own numbers rather than recomputing alignment) gives a curve that attenuates with k: concordance is -0.573 at $k = 12$, -0.574 at $k = 10$, -0.556 at $k = 5$, -0.490 at $k = 2$ and -0.393 at $k = 1$ (Spearman $\rho = -0.90$ against k; the ordering is monotone except that $k = 10$ and $k = 12$ are indistinguishable at -0.574 against -0.573). What the curve does *not* do is approach the values on the cohorts that fail to reverse: at $k = 1$, the per-subject depth of Brennan and MEG-MASC, 92% of draws are still negative and the mean sits at -0.393 against Brennan's $+0.37$ and MEG-MASC's $+0.45$ (re-extracted; $+0.34$ and $+0.49$ before). The dispersion tells the same story from the other side, falling from ± 0.181 at $k = 1$ to ± 0.032 at $k = 10$: removing sessions makes the estimate noisier without moving its centre.

The strongest form of this check uses no averaging at all. Scoring each session on its own, 11 of the 12 are negative (values -0.63 to $+0.03$, mean -0.389, median -0.421). We report this descriptively and attach no p-value: the twelve sessions share one listener, and Appendix K declines to pool exactly such units for exactly this reason, a standard that must also apply where the result is in our favor. Each session is nonetheless scored at a depth matching the cohorts where the reversal does not appear.

Depth therefore governs the precision of the concordance estimate and not its sign, and the difference between cohorts has to be sought in modality, listener sample or stimulus instead. The objection that the reversal is visible only because this cohort supplies unusually much data per subject is therefore not supported by the depth curve (`results/p4_encoding/concordance_depth_curve_CORRECTED.json`). Depth is not the only difference the cross-cohort comparison crosses. On LibriBrain encoding is scored on a held-out session, that is, on text the ridge map never saw; on MEG-MASC and Brennan it is scored on a held-out listener who heard the same stories as the training listeners (`p4_concordance32.py`, leave-one-group-out by session or by subject). This fold difference, like the tokenization rule of Appendix D, separates the reversing cohort from the others and is not tested here.

J Static-Embedding Control (C4)

The tensor route partials out a fourth confound, C4, the dissimilarity matrix of the model's own static layer-0 embedding, which removes the part of a model's contextual geometry that its non-contextual token embedding already predicts. The fixed-depth route of Section 3 did not. We added it to the 32-checkpoint LibriBrain comparison under a pre-registration committed before any C4 score existed (`PREREG_c4_static32_20260924.md` in the supplementary material). For every checkpoint, C4 is the cosine-distance matrix of its layer-0 hidden state read at the same target token(s), from the same extraction as its scored depth-0.5 features, on the same 30,000 words and 12 sessions; Eq. (1) gains C_4 as a fourth conditioning matrix and the encoding scores are unchanged. The primary is the Spearman concordance of the session-mean C1–C4 similarity scores with the encoding scores over the 32 checkpoints, judged by its family-clustered interval: the reversal holds if the interval lies below zero, is not determined if it contains zero with a negative point estimate, and is removed if the point estimate is non-negative; a split-half reliability of the C1–C4 ranking below 0.70 would have made the result indeterminate.

Gates. Recomputing the C1–C3 scores in the same run reproduced all 32×12 stored values to 7×10^{-17} and the concordance -0.5733. Re-extracting the depth-0.5 features reproduced the cached ones for 31 checkpoints (median row cosine above $1 - 10^{-10}$, every row above 0.9999) and GPT-2 small passed by file identity (Deviation 1 below), so each C4 comes from the same model and tokenization as the scores it controls. No re-extraction fell back from a target token, and no C4 entry was non-finite.

Result. Table 5 gives the numbers. The concordance moves from -0.573 to -0.415. Its checkpoint-level Fisher interval still excludes zero, but the registered family-clustered interval reaches $+0.006$, so under the pre-registered rule the reversal is *not determined* once C4 is partialled out. The change itself is not determined either: the paired interval of Δ contains zero, so the data do not show that C4 weakens the reversal. The C1–C4 ranking stays reliable (0.97), so the open interval reflects the concordance, not a noisier similarity ranking. C4 does not standardize or whiten features, so it does not test the geometry account of Section 5.3, and because each C4 comes from the same extraction as the features it controls, it does not test the extraction-rule difference of Appendix D. C4 was not run for the 57-checkpoint set, MEG-MASC or Brennan.

Class (n)	mean RSA, C1–C3	mean RSA, C1–C4	mean encoding
Autoregressive (18)	0.0049	0.0044	0.0494
Masked (9)	0.0082	0.0060	0.0412
Encoder–decoder (5)	0.0101	0.0087	0.0426
	C1–C3	C1–C4	Δ (C4 − C3)
Concordance with encoding, ρ	-0.573	-0.415	$+0.159$
Family-clustered 95% CI	$[-0.79, -0.26]$	$[-0.73, 0.01]$	$[-0.03, 0.42]$ (paired)
Fisher-z 95% CI	$[-0.77, -0.28]$	$[-0.67, -0.08]$	—
Split-half ranking reliability	0.98	0.97	—

Table 5: The 32-checkpoint LibriBrain readout comparison with and without the static-embedding control C4. Family-clustered intervals resample the 12 model families (10,000 draws, seed 0); Δ's interval uses the same draws for both concordances. Receipts: `p4_encoding/concordance32_c4.json` (aggregate) and `p4_encoding/c4_static32/` (one gate-and-score receipt per checkpoint).

Deviations from the pre-registration. Both were decided after the gates were read and before any C4 value was. (1) GPT-2 small: its cached features come from the older single-model extractor, so re-extraction could not pass the identity gate; its C4 is the layer-0 file of that same extraction, checked by file identity. (2) The other 17 autoregressive checkpoints failed the identity gate when re-extracted with the documented extractor, because their cached features were written by an earlier revision (Appendix D). They were re-extracted with that revision verbatim, with the gate unchanged, and all 17 then passed. To finish before the deadline these 17 extractions ran on CPU rather than GPU; the unchanged gate compares them with the GPU-extracted caches.

K Broadband versus Narrowband (30–80 Hz) Gamma

The benchmark also speaks to a preprocessing convention. Intracranial language pipelines filter each electrode to the high-gamma envelope, a convention established in intracranial electrophysiology (Crone et al., 2006; Ray & Maunsell, 2011), standard in speech decoding (Anumanchipalli et al., 2019; Chen et al., 2024) and in the ECoG language-model literature (Goldstein et al., 2022), before comparing with a model, and the logic carries over informally to non-invasive work. Because the band is a benchmarked dimension here, we can compare two available band definitions within this protocol: for each recording unit, one (session or patient) $\times$ model pairing, we take the peak broadband alignment over the region, time, and layer grid and the peak gamma-band alignment over the same grid, and compare them. One scope note applies: our gamma band is 30–80 Hz (Appendix D), the band the benchmark computes for every cohort, which is lower than the 70–150 Hz high-gamma envelope some intracranial pipelines use, so the comparison is against the gamma-band operationalization, not that exact envelope. That restriction is structural rather than a choice we could revisit: the MEG cohorts are sampled at 250 Hz, so frequencies above 125 Hz are not representable in them, and the one cohort where the convention actually applies is the intracranial one, for which we hold derived alignment tensors rather than the epoched recordings a re-banding would require. Testing the envelope directly therefore means re-deriving the ECoG tensors from source, which is the clearest extension of this result.

We report the band comparison at the level of the recording, averaging each session or patient over the models scored on it, because the same recording contributes one row per model and pooling those

rows would count it three times. Across the four band-resolved cohorts (36 LibriBrain recordings, 76 MEG-MASC, 92 MEG-SCANS, 9 ECoG patients; 213 recordings) broadband reaches the higher peak in 206, 96.7% (Figure 6), and the direction holds in every cohort and does not depend on modality: LibriBrain $36/36$, ECoG $9/9$, MEG-SCANS $88/92$, MEG-MASC $73/76$. Recordings still nest within listeners, and LibriBrain's 36 recordings come from a single listener, so we do not pool these into one sign test: the pooled p would be anticonservative by orders of magnitude and we do not report it. The one test here over unambiguously independent subjects is ECoG, where broadband wins in 9 of 9 patients (one-sided sign test $p = 0.002$); the cohort means are also ordered the same way in all four cohorts ($4/4$, one-sided $p = 0.0625$ treating the cohort as the unit, the most conservative reading available). Three MEG systems, one intracranial system, English and German, autoregressive and masked models, and in each slice the broadband peak sits above the gamma-band peak for the large majority of units.

The reason is reliability, and this is where the benchmark's built-in ceiling earns its place. On LibriBrain at the temporal_L, 350 ms cell (462 shared words, 12 sessions) the per-band split-half ceiling (Table 6) shows that gamma (30–80 Hz) carries no reproducible word-level structure at all: its split-half word-dissimilarity reliability is -0.002, indistinguishable from zero. The low-frequency, high-power bands do reproduce, with ceilings of 0.135 for broadband, 0.244 for delta, and 0.107 for theta, while alpha (0.003) and beta (0.001) sit at zero with gamma. Raw alignment tracks the ceiling: broadband 0.0056, delta 0.0055 (a near tie), theta 0.0016, gamma -0.0003. Gamma loses because it has no reliable word-level geometry to align to, not because a mid-frequency carrier is out-competed.

That mechanism is cohort-specific, and we checked rather than assumed it. Repeating the per-band reliability on Brennan, split half the listeners against the other half because no single scalp-EEG subject has the repetitions for a stable matrix, the ordering *reverses*: gamma is the most reproducible band ($+0.052$) and delta and theta sit at zero (-0.002, -0.002). A plausible reading is that scalp gamma is dominated by stimulus-locked muscle activity, which is reproducible across listeners hearing the same story without being word representation, but we have not established that. What follows for the claim above is narrower than it first appears: the band comparison rests on the four cohorts that carry the full protocol, which do not include this EEG probe, and the reliability explanation is demonstrated on LibriBrain rather than shown to be general (`results/megmasc_meta/perband_brennan_across.json`).

We keep the honest limit attached, because the reliability mechanism is real but it does not license a spectral-carrier claim. Two controls dissolve the stronger reading. Ceiling normalization is undefined for the band the claim contrasts against: dividing each band's alignment by its own ceiling gives broadband a normalized value of 0.04 and gamma 0.14, but gamma's ceiling is ≈ 0, so its normalized value is a tiny alignment over a tiny ceiling, an uninterpretable quotient rather than a reversal. Confound control removes the group-level advantage: after partialling out word length, frequency, orthography, and the static embedding, broadband's partial alignment goes to -0.0036 and no band with a positive ceiling retains positive partial alignment at the group mean (beta's $+0.0006$ sits on a ceiling of ≈ 0). So the defensible claim is narrow and it is a preprocessing correction. Under the tested definitions, broadband preserves more reproducible word-level structure on LibriBrain (the field's 70–150 Hz high-gamma envelope is not directly tested here); do not read the effect as evidence that broadband oscillatory power is the neural code of alignment. The per-unit dominance is the load-bearing result and it is stable; the group-level frequency reading is weaker than the per-unit summary alone would imply.

Group-level dissolution of the broadband effect. The per-unit broadband advantage is stable, but the group-mean frequency reading does not survive ceiling normalization or the confound battery. The broadband advantage is a per-unit preprocessing correction (93.1%, $595/639$ unit$\times$model rows; equivalently 96.7%, $206/213$ at the recording level), not proof of a broadband neural carrier.

ECoG coverage. The strongest ECoG cells land on precentral gyrus rather than superior temporal gyrus, most plausibly because of where the dense grids were placed. The ECoG numbers support the direction and magnitude of the broadband advantage, not a specific cortical localization.

Band	Hz	split-half ceiling	raw alignment	partial (confound-controlled)
gamma	30–80	−0.002	−0.0003	−0.0006
broadband	1–80	0.135	0.0056	−0.0036
delta	1–4	0.244	0.0055	−0.0064
theta	4–8	0.107	0.0016	−0.0057
alpha	8–13	0.003	−0.0034	−0.0021
beta	13–30	0.001	−0.0001	0.0006

Table 6: Per-band split-half reliability ceiling (Spearman-Brown corrected) and alignment on LibriBrain (temporal_L, 350 ms, GPT-2 layer 5, 12 sessions, 462 shared words). Only the low-frequency, high-power bands carry a positive, reliably measured ceiling and positive raw alignment; gamma's ceiling is ≈ 0 so its alignment cannot be normalized. At the group mean, no band with a reliably positive ceiling keeps positive partial alignment, so the raw broadband advantage here is largely lexical-confound-driven. Source: `results/p2_perband_ceiling.json`.

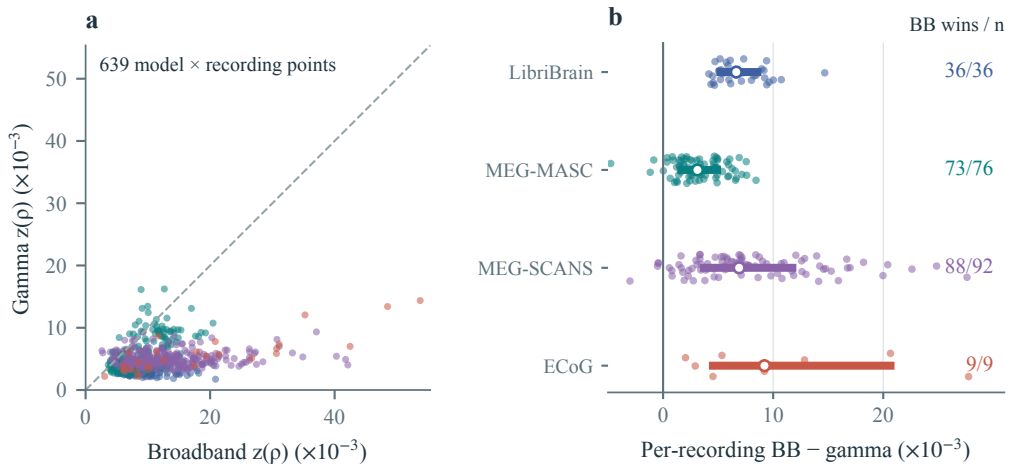


Figure 6: **Broadband advantage at the recording level.** (a) All 639 model-by-recording pairs, colored by cohort; points below the identity line favor broadband. These are repeated measurements, not 639 independent units. (b) After averaging the three models within each recording, all 213 recording-level differences are shown; thick segments are the interquartile range, white dots the median. Right labels count positive differences. Broadband wins in 206/213 recordings. Source: per_unit_bb_vs_gamma.tsv.

L Literature-Inspired Hypotheses Under Changed Settings

This analysis transports ten literature-inspired hypotheses into the benchmark's settings (Table 8; detail in Appendix P). It mixes conceptual extensions and analogies, often changing modality, task, model or readout. It is therefore not a direct replication campaign and its labels cannot estimate a literature replication rate. The bundled `scorecard_transport.tsv` separates source and tested estimands, changed settings and evidence availability. Historical *replicate/refine/contradict* labels remain in the original files for provenance; the table here uses scoped outcomes and marks the baseline-comparison row unverified because its source tensors are missing. Figure 9 is a companion matrix of our own findings by cohort, not a tally of which published papers were correct.

One conceptual analogue supports its directional prediction: the shared and experience-dependent encoding tiers of Bhaya-Grossman et al. (2025) — reported *within the same STG neural populations*, for word boundaries, word frequency and language-specific sound sequence statistics — have a layer-depth analogue on MEG-SCANS, where multilingual-BERT wins at shallow layers and German-BERT wins at deep layers.

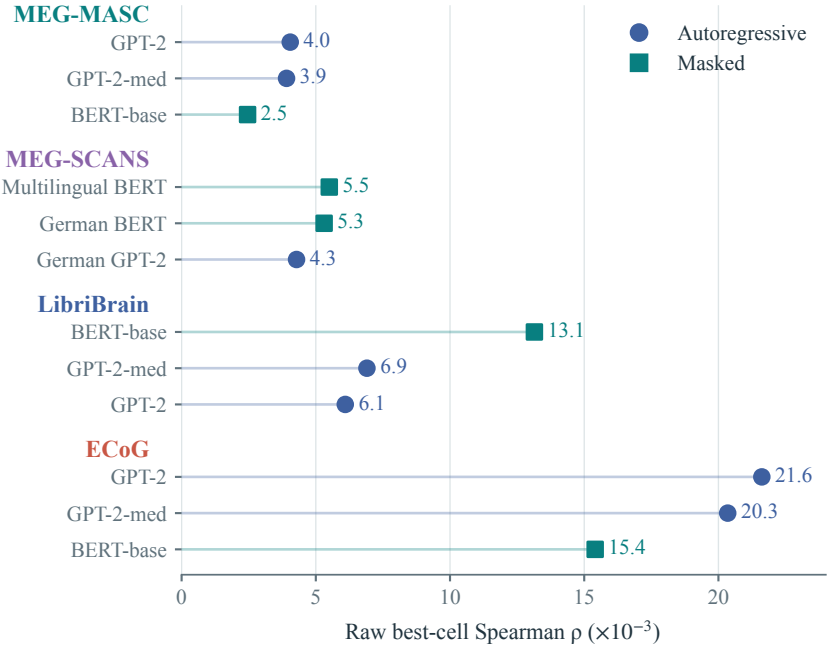


Figure 7: **Raw best-cell alignment varies by cohort and architecture.** Each named point is read directly from cross_cohort_summary.tsv; square markers denote masked models and circles autoregressive models. Values are raw Spearman correlations, not the confound-controlled primary score. German models in MEG-SCANS are compared only within that cohort. Cohort means remain 3.5, 5.0, 8.7, and 19.1 in units of 10^{-3}.

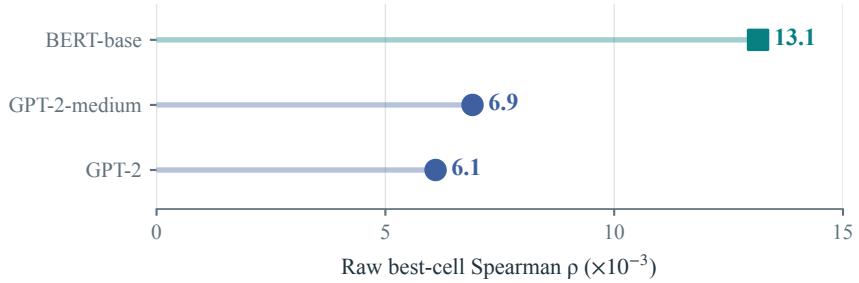


Figure 8: **LibriBrain best-cell leaderboard for the three source-backed models.** Values are read directly from cross_cohort_summary.tsv (raw Spearman correlation, best band per model): BERT-base 13.1, GPT-2-medium 6.9, GPT-2 6.1, in units of 10^{-3}. The old Random Gaussian mark is omitted because its source was not recomputable in the supplied receipts; it is not used as a verified baseline here.

The mixed outcomes describe support under some tested settings but not others. Caucheteux et al. (2023) report that augmenting a model's present-word activations with predictions spanning multiple timescales improves the mapping to fMRI responses in 304 listeners. Under a lag analogue on MEG the anticipation holds only in a window: shifting the model's word sequence against the brain's, the autoregressive model's alignment peaks one word *ahead* in the 350–550 ms windows, but peaks *behind* at 150–250 ms and at zero lag when collapsed over time, and the bidirectional model shows no forward peak at any window (`results/exp_caucheteux/time_resolved_kpeak_libribrain.tsv`). Anticipation is present, but confined to one model

and one latency range. The mid-layer preference of Toneva & Wehbe (2019) replicates on all four MEG/EEG cohorts in its weaker form, mid-to-shallow over deep; as they state it — middle layers beating both shallower and deeper — it holds in two of the eight recomputable cells. It inverts on ECoG, where deep layers win. The contextual-over-static advantage underlying Schrimpf et al. (2021)'s next-word-prediction correlation, which we do not re-measure, holds in breadth (contextual models beat GloVe and a phonemic baseline in cell count and layer-depth gradient) but not in single-cell magnitude, where a historical, unverified comparison had GloVe-300 (Pennington et al., 2014) tying GPT-2 (6.7 against 6.1×10^{-3}; see Evidence status below). Goldstein et al. (2025) report the layer–peak-time hierarchy for two causal models, GPT-2 XL and Llama-2, in ECoG high-gamma and strongest in Broca's area. Restated across cohorts and bands it reproduces for BERT temporal_R ($r = +0.73$ broadband), an architecture they did not test, and inverts for the GPT family in the delta band. The constituent-boundary effect of Zou et al. (2026) holds in 7 of 12 English cells, and in German it is model-dependent rather than uniformly reversed: on MEG-SCANS multilingual BERT reverses significantly ($7/23$, $p = 0.005$) while native German-BERT is null ($p = 0.96$) and German-GPT-2 only trends negative ($p = 0.09$). The verdict here is the scorecard's own operationalization, computed from the same cohorts. The decoding success of Tang et al. (2023) leaves a weaker relative in trial-level MEG, in a task family they explicitly separate from their own: contrastive retrieval clears chance on all seven targets (six models, GPT-2 at two segmentations; 5.7–6.5× at top-10, shuffled-target control 0.7×), yet absolute accuracy stays far below continuous reconstruction (top-1 0.03%, median rank 2143 of 9495), so what transfers is decodability, not continuous language. The confound critique of Hadidi et al. (2026) — positional signals and word rate — is real but does not kill the signal under disaggregation (per-subject MEG-MASC GPT-2 retains 36% in 6 of 10 subjects). He et al. (2025) disentangle lexicon, syntax, meaning and reasoning, on one model family; the medial-minus-onset contrast of their reasoning residual sign-inverts between GPT and BERT here, so the architecture condition is ours, not theirs, and the meaning residual inverts the other way and far more weakly.

The size-ordering extension is not supported under our encoding protocol. Antonello et al. (2023) report that fMRI encoding performance scales logarithmically with model size from 125M to 30B for unidirectional models (OPT and LLaMA families). Restated under our encoding readout, it does not hold. One observation sits outside their model class and we report it as such: on raw headline alignment BERT-base (110M, masked) leads LibriBrain at 13.1×10^{-3}, above the larger autoregressive GPT-2-medium (355M, 6.9×10^{-3}) and GPT-2 (124M, 6.1×10^{-3}; `cross_cohort_summary.tsv`, Figure 8), so a small masked model outranks a three-times-larger causal one.

The size ladder itself does not climb. Scored per session at the pre-registered depth, so that no best-layer selection enters, the six-point Pythia ladder from 70M to 2.8B does not order alignment by size ($\rho = -0.14$ across the six sizes, $p = 0.79$), and the largest model sits nominally *below* the smallest despite forty times the parameters (paired within session, 2.8B minus 70M $= -2.1 \times 10^{-3}$, Wilcoxon $p = 0.23$; `scaling_ladder_test_CORRECTED.json`). Under the corrected encoding arm the non-ordering is, if anything, sharper than under best-layer scoring. Alignment rises modestly through 410M–1B and comes back down by 2.8B rather than plateauing. One detail changes relative to best-layer scoring: a three-point version of this ladder, scored at each model's best layer, appeared to peak at 160M, but that was best-layer selection — at the fixed cell 410M significantly exceeds 160M ($p = 0.005$).

One limit bounds the verdict: the sharpest size comparison (BERT-base versus GPT-2-medium) confounds parameter count with objective and tokenizer. The registered modern extension (§5.4, Figure 4) removes the range limit, but its Qwen2.5 encoding arm used RidgeCV alphas 1e1–1e6; the original-grid ceiling is selected in $11/12$ sessions at every re-extracted rung from 3B to 72B (`space_fix_20260925.json`, per-rung score receipts), as an earlier three-rung probe found (`alpha_probe.json`; its 32B and 72B rungs on the unspaced features). The same bounded grid scored the encoding readout of the 32-checkpoint comparison, which includes Qwen2.5-0.5B and Qwen2.5-1.5B; at 0.5B the grid top is selected in 11 of 12 folds (`alpha_probe.json`), the selected alphas of the other checkpoints are not recorded in the concordance receipt, and the sensitivity of the readout concordance to the ridge range is untested. The seven-rung post hoc replay in Table 7 retains all 12 outer leave-one-session-out folds, the same 30,000 words, depth and neural target; it reproduces every original-grid session score (maximum absolute discrepancy 2.2×10^{-16}), then selects from the expanded 1e1–1e14 grid using training data only. On the re-extracted rungs the size correlation weakens from -0.964 to -0.464 (exploratory $p = 0.294$ on seven rungs; -0.893 on the

Table 7: **Post hoc training-only alpha-grid sensitivity on the seven Qwen2.5 rungs.** Mean held-out encoding correlation over the same 12 LibriBrain sessions. Original alpha grid: logspace$(1, 6, 8)$; expanded grid adds 1e7, …, 1e14. The mode alpha is selected in $11/12$ folds at every rung except 14B $(10/12)$; the other sessions select an original-grid interior value or an adjacent grid value. Rows 3B–72B are re-extracted after the tokenization defect (Appendix D); on the unspaced features they read $0.05951/0.06105$, $0.05641/0.05453$, $0.05252/0.05385$, $0.04856/0.05520$ and $0.03325/0.04491$. The registered Figure 4 shows the original column. This table is a sensitivity analysis, not a new registered scaling test.

Qwen2.5	Width	Original R	Expanded R	Mode α
0.5B	896	0.05933	0.05550	1e7
1.5B	1536	0.05703	0.05079	1e7
3B	2048	0.05806	0.06011	1e7
7B	3584	0.05669	0.05348	1e7
14B	5120	0.05164	0.05326	1e8
32B	5120	0.04926	0.05481	1e8
72B	8192	0.03195	0.04759	1e14 (boundary)

unspaced features): the width-matched 14B/32B contrast and the 32B–0.5B gap are sensitive to the regularization range. 72B still scores below 0.5B by 0.0079 on the expanded grid (0.0106 on the unspaced features), so the extension does not establish a flat size relation.

At 72B the training-selected alpha reaches the new grid ceiling in $11/12$ folds, so it is not an interior optimum. The rest of this paragraph (the fixed-alpha diagnostic and the three quantities it separates) was computed on the unspaced 72B features before re-extraction (Appendix D) and was not rerun; its training-selected 0.0449093 is the unspaced counterpart of the 0.04759 in Table 7. A separate fixed-alpha diagnostic compares 1e10, 1e12 and 1e14 with the analytic $\alpha \to \infty$ prediction direction: their scale-free mean correlations are $0.0458118/0.0458094/0.0458094$ versus the limit 0.0458094, with maximum per-session deviation 5.8×10^{-10} at 1e14. This certifies a large-penalty *direction* plateau, not a finite optimum. Keep three quantities distinct: the training-selected original scorer gives 0.0449093 (11 folds at 1e14, one at 37275.94); removing the fixed denominator epsilon at those same selected alphas gives 0.0450517. Fixing 1e14 in *all* folds gives 0.0456666 with the original scorer and 0.0458094 without the epsilon. Only the same-policy differences (0.0001424 or 0.0001429) measure epsilon attenuation; the larger cross-policy gap also changes the exceptional fold's alpha.

A registered second-cohort ladder removes the single-listener limit: the same seven Qwen2.5 rungs, extracted and scored on MEG-MASC's merged 27 subjects at the published MASC cell with subject replacing session as the cluster unit, both cross-cohort readings committed before extraction, and all seven re-extracted after the tokenization defect (Appendix D). The outcome splits by readout. The similarity leaderboard passes the registered rankability gate (split-half 0.83, 90% CI $[0.13, 0.98]$) and size does not order it: $\rho = +0.29$ with a subject-clustered CI $[-0.32, +0.75]$, and 72B minus 0.5B paired within subject is $+0.9 \times 10^{-4}$ ($p = 0.71$, $15/27$ positive). The encoding leaderboard has split-half rank reliability 0.70 with a 90% CI $[-0.16, 0.94]$ that includes zero, so it fails the gate and that arm is indeterminate — not folded into either reading, per the registration — and no cross-cohort encoding statement is made. The gate's criterion, a 90% split-half CI excluding zero, was fixed in the registration before any checkpoint above 1.5B had been extracted on the MASC word list, and was applied identically to both arms. On the unspaced features the gate outcomes were the reverse: encoding passed (0.81, $[0.40, 0.96]$) with $\rho = +0.29$, CI $[-0.43, +0.64]$, and similarity failed (0.76, $[0.00, 0.96]$) (`adr0010_q1_masc_verdicts.json`; re-extracted: `adr0010_q1_masc_verdicts_spacefix.json`).

The closest opposing result is worth stating precisely, because it runs on one of our own cohorts. Hong et al. (2026) score four model families from 82M to 70B against the same intracranial recordings we use as our ECoG cohort (`ds005574`) and report that larger models predict neural activity better, logarithmically in size. Three differences separate that result from ours rather than adjudicating between them. Their encoding score is the maximum correlation over all lags *and* all layers per electrode, so it inherits exactly the best-cell selection our held-out split prices at 6–85% (§4.3), whereas our ladders are scored at a depth fixed before extraction. Their comparison is encoding-only,

Source hypothesis	Benchmark extension	Outcome
Caucheteux et al. (2023)	predictive-distance lag on LibriBrain	mixed
Bhaya-Grossman et al. (2025)	mBERT/German-BERT layer split, MEG-SCANS	analogue supported
Toneva & Wehbe (2019)	mid-layer scan, 5 cohorts	mixed (weak reading; ECoG inverts)
Schrimpf et al. (2021)	LLMs vs GloVe/phonemic, 5 cohorts	unverified (tensors missing)
Goldstein et al. (2025)	layer $\times$ peaktime, 5 cohorts	mixed (band-conditional)
Zou et al. (2026)	boundary RSA, 4 cohorts $\times$ 6 LLMs	mixed (German model-dependent)
Hadidi et al. (2026)	confound battery, per-subject	mixed (survives disaggregation)
He et al. (2025)	reasoning residual, LibriBrain	mixed (boundary split architecture-dependent)
Tang et al. (2023)	trial-level MEG decoding	mixed (transfers weakly)
Antonello et al. (2023)	Qwen + Pythia encoding ladders	not supported here (encoding does not rise; slope regularization-sensitive)

Table 8: Cross-setting tests, not a replication-rate estimate. Outcomes apply to the listed benchmark extension. Rows marked "5 cohorts" include the uncontrolled Brennan EEG probe; the source-estimand and availability inventory is `scorecard_transport.tsv`.

while our claim is that the two readouts order size in opposite directions on the largest ladder. And their own curve is not monotone at the top: they report a plateau near 7B and a decline at 66B, which is the saturation our registered verdict states rather than a contradiction of it. What their result does establish, and ours does not dispute, is that under a single readout with per-electrode selection, size buys predictive accuracy on this cohort.

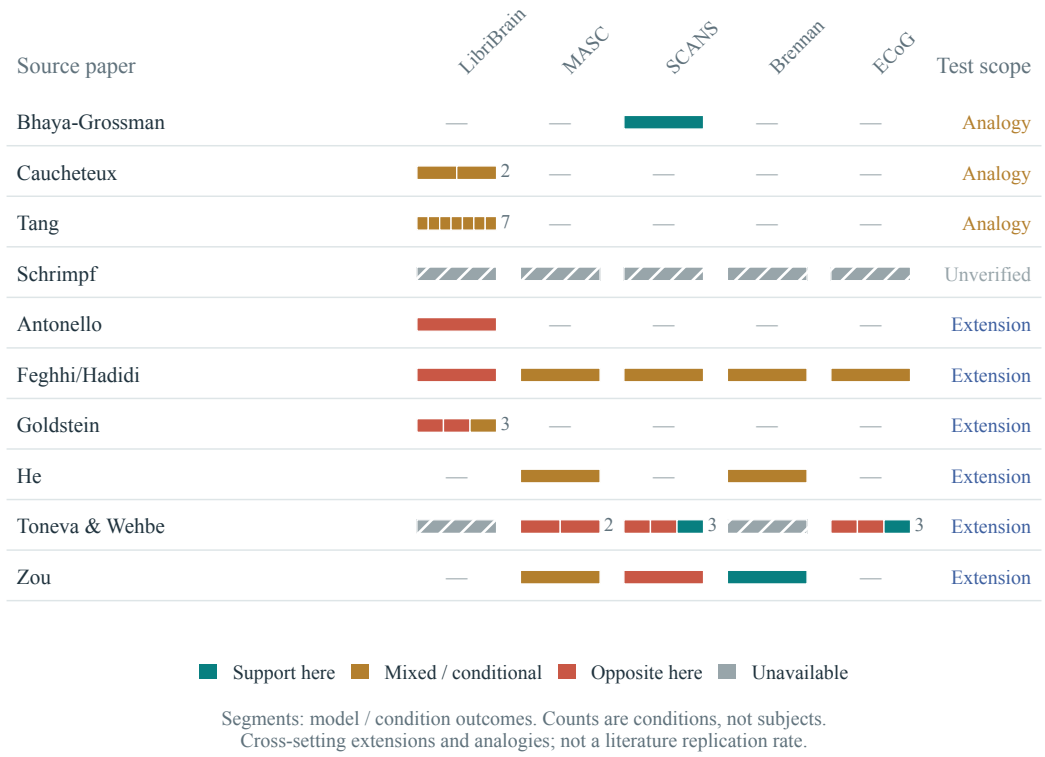


Figure 9: **Cross-setting tests and source coverage.** The right column follows scorecard_transport.tsv: these tests are conceptual extensions or analogies, not literal replications. Each mini-bar displays the stored per-model/condition outcomes for that cohort in scorecard_derived.json. Colors describe support, conditional/mixed behavior, or the opposite direction under this benchmark's operationalization, not verdicts on the original papers. Numbers count conditions, not independent subjects. Dashes mean no stored entry; hatched gray denotes unavailable evidence. The Schrimpf comparison remains unverified because tensors are missing. No replication fraction is computed.

Evidence status. Nine analyses have executable output-to-outcome mappings; the baseline comparison does not. Its per-model LibriBrain tensors are missing, and the surviving GPT-2 maximum is selected over cells where both GPT-2 and BERT passed a threshold, while GloVe's maximum is unconstrained. That asymmetry favors the baseline, so the reported tie is historical and unverified. It does not enter a replication tally. Other mappings also changed during audit: the lag test is an analogue of forecasting, the layer-split result is a between-model contrast, and the decoding result concerns retrieval rather than continuous reconstruction. These changes are retained in the provenance files. In particular, the multilingual layer-depth crossover is an analogy on one cohort, not a replication of the original within-STG physiological contrast.

M Narrative Identity versus Finite Word Sampling

Why the registered primary and co-primary disagree. The registered primary of ADR 0005 was the variance of the leaderboard across the four stories over its variance under participant resampling, with confirmation requiring all three criteria jointly; it came out 0.80 (similarity) and 0.16 (encoding) against a threshold of > 2, so the test was not confirmed and falsifier F1 fired (§5.2). Post hoc inspection shows why the co-primary nonetheless met its criterion: all four story-specific rankings are displaced from the pooled ranking by a similar amount ($\tau = 0.29$–0.44, against 0.87 for a participant resample), so they are uniformly far from it rather than dispersed around one another, and a variance ratio cannot see a uniform displacement. That explanation is post hoc — the failure

mode was not registered — so it cannot restore confirmatory status. A confirmatory version requires a new registration with a statistic sensitive to displacement, on a new stimulus set.

Numbers behind Section 5.2. At full length each story's similarity leaderboard has Spearman–Brown split-half reliability 0.933–0.958, and the six between-story agreements have median -0.411 and mean -0.025. Capped at the shortest story (F4) they have median -0.146 and mean $+0.003$, and The Black Willow's reliability falls to 0.47. On eight Brennan segments of about $8\,000$ words, similarity (Spearman–Brown) reliability is 0.995–0.999, while encoding is undetermined on seven.

Words and narratives vary together above, so a further split separates them. Each story is divided into two equal, disjoint word partitions, giving eight leaderboards in which a within-story pair shares the narrative and differs in words, while a between-story pair differs in both, at matched size throughout.

The two readouts behave differently. Similarity leaderboards from the same story agree at $+0.556$ on average and from different stories at -0.008, a difference of $+0.564$ (95% CI $[+0.35, +0.78]$, one-sided $p = 0.012$): a suggestive within-story advantage that does not establish a narrative-identity mechanism, given the failed rankability condition below. Encoding shows no such effect — $+0.509$ within against $+0.476$ between, difference $+0.034$, CI $[-0.03, +0.09]$ — so encoding rankings generalize across narratives about as well as they generalize across disjoint halves of one.

This analysis does not meet our pre-registered rankability precondition and is reported as suggestive: at $\approx$ $16\,000$ word-trials per partition only one of eight partitions has both leaderboards individually determined. Two things argue the signal is real anyway — within- and between-story pairs use the *same* partitions with the same reliabilities, so attenuation would shrink both rather than open a gap, and within-story agreement tracks partition reliability in the direction attenuation predicts ($0.86/0.64 \to +0.74$ against $0.14/0.32 \to +0.36$). Settling it needs longer narratives than this cohort provides.

N Three Numerical Failure Modes and Their Detection

This appendix reports in full the computational failures summarized in §5.5. All three share one cause: code written as though its inputs were on a unit scale, applied to neural data stored in physical units, and all three produced plausible numbers rather than errors.

Channel-type metadata. Trusted channel-type metadata let 14 EKG and DC channels carrying 99.97% of total variance into an intracranial analysis, making the result a property of the stimulus audio track rather than of cortex. Caught by a per-channel variance audit, and fixed by validating channel types against the recording's own variance profile.

Ridge grid against a rebuilt archive. A fixed ridge grid met a rebuilt archive at std $\approx 10^{-13}$ and returned encoding correlations of exactly 0.00000 for all 32 models — diagnostic enough to catch on inspection, unlike the third failure.

A guard constant in the correlation denominator. A 10^{-12} guard in a correlation denominator: at std $= 3.95 \times 10^{-7}$ the denominator is of order 10^{-11}–10^{-10}, and the guard sits beside the variance of *that model's own* predictions, so a stronger-predicting model is shrunk less. It reordered models scoring between $+0.029$ and $+0.051$, inflating the readout concordance of an earlier analysis in this project from -0.573 to -0.678; changing only that constant recovers -0.573, matching an independent scale-invariant fix to sixteen significant figures. The claim-reproduction harness read $97/97$ green before and after (the registry has since grown to 193 claims), because it checks stored results against the paper's claims and cannot see how the results were produced.

What the bundle holds. The original artifact, the corrected one, and the isolation experiment that changes only the guard constant, with harness checks that pin all three values. These are stored-value checks: the harness does not yet include a test that rescales the neural data, which is what would detect this failure class before it produces a number.

Table 9: **Post hoc rank-one sensitivity, both readouts recomputed on the same ten models.** Cells are RSA/encoding ranks (rank 1 = highest) at the fixed LibriBrain primary cell. Center-only is $X - X_{\text{mean}}$; doses $0.3/0.7$ add a sample-varying rank-one term to that centered matrix. The old frozen-encoding RSA-only injection is not mixed into these ranks. Source: `rogue_encoding_20260922/summary.json` and ten four-condition model receipts.

Model	Raw	Center only	Rogue 0.3	Rogue 0.7
Pythia-70M	7/2	6/2	6/1	5/1
Pythia-160M	5/3	7/3	7/8	7/8
Pythia-410M	9/1	9/1	9/7	9/7
GPT-2	8/8	8/8	10/5	10/6
GPT-2-medium	10/6	10/6	8/2	8/2
BERT-base	4/7	5/7	5/4	6/4
BERT-large	6/9	4/9	4/9	4/9
RoBERTa-base	3/4	3/4	3/3	3/3
T5-base	2/10	1/10	2/10	1/10
T5-large	1/5	2/5	1/6	2/5

O Geometry Sensitivity and Unresolved Explanations

Section 5.3 established that the reversal is a class-level offset (removing the architecture-class means leaves a within-class concordance of -0.25, n.s., and the encoding-minus-similarity rank gap separates causal models from the rest by $+25.7$ ranks, family-clustered two-sided $p = 0.005$), and that standardizing the features attenuates it to a residual whose family-clustered CI crosses zero. With width matched, the gap stays $+14.1$ ranks over 22 checkpoints (family-clustered two-sided $p = 0.018$); on Brennan EEG the class gap is -0.4 ranks ($p = 0.53$; computed before its three unspaced checkpoints were re-extracted). This appendix reports the candidate mechanism we tested and could not support, since a benchmark paper should be as explicit about a failed explanation as about a confirmed one.

RSA-only sensitivity, not an encoding-invariance test. We transformed the embeddings of ten models by either adding a random sample-varying rank-one component or rescaling coordinates. The resulting similarity ranking changes: its correlation with the original ranking falls to 0.49 at the largest rogue dose, and to 0.76 under coordinate rescaling. The historical analysis correlated these new RSA scores with a *frozen*, previously computed encoding ranking, obtaining -0.37 to -0.49 around -0.43 (`invariance_intervention.json`). It did not refit encoding after intervention; these are historical RSA-only sensitivity numbers, not evidence of encoding invariance.

The claimed invariance behind that design was wrong for the rogue arm. Train-fold standardization removes coordinate scale changes, but adding $sav^\top$, where a varies across samples, changes the standardized design. We therefore ran a separate *post hoc* primary-cell diagnostic: the same ten models, $30{,}000$ shared words, 12 LibriBrain sessions, depth 0.5, and the original outer-session ridge protocol, with both RSA and encoding actually recomputed at raw, center-only, and centered-plus-rank-one doses $0.3/0.7$. All ten raw baselines replay their stored per-session scores within 1.1×10^{-16}. Centering alone changes RSA order ($\rho = 0.927$ vs. raw) but leaves encoding scores unchanged to 2.1×10^{-17} and encoding ranks identical. After rank-one injection, encoding rank correlation with raw falls to $0.406/0.442$ (largest score shifts $0.0189/0.0255$); the standardized-ridge invariance premise is falsified on real data (Table 9).

The RSA–encoding rank concordance changes from -0.261 raw to -0.406 centered and $-0.248/-0.164$ after the two added-direction doses. Five-family block-bootstrap intervals for the encoding-rank correlations at doses $0.3/0.7$ are wide ($[-0.098, 1.000]/[-0.105, 1.000]$); the ten models do not identify a general mechanism. The perturbation is synthetic and is not the models' endogenous anisotropy. Thus actual encoding movement overturns the old invariance premise, but neither proves that anisotropy causes the original cross-readout disagreement nor excludes other geometry explanations.

Anisotropy. Cosine-distance RSA is sensitive to a dominant mean/rogue direction, while ridge encoding standardizes its inputs. Autoregressive models are extremely anisotropic: the median causal

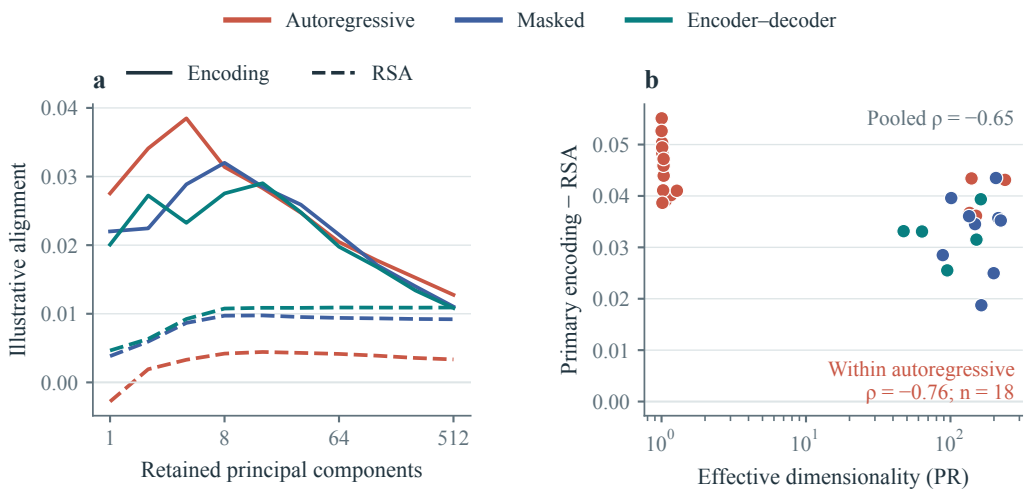


Figure 10: **Effective dimensionality covaries with the readout gap.** (a) Architecture-averaged PCA-truncation curves, shown as an illustrative diagnostic; solid is encoding, dashed RSA. (b) Participation ratio against the corrected primary-score gap for all 32 checkpoints. The within-autoregressive relation (rho=-0.76, n=18 in six families) is highlighted; the pooled rho=-0.65 is not interpreted causally or as 32 independent models. Family-clustered intervals and the between-class decomposition remain in the text. Sources: pctrunc receipts and concordance32_CORRECTED.json.

model has participation ratio 1.0, its top principal component alone explaining $\approx$ 98% of feature variance, the known anisotropy and "rogue dimension" phenomenon of transformer language models, most extreme in autoregressive ones (Ethayarajh, 2019; Timkey & van Schijndel, 2021; Gao et al., 2019); masked and encoder–decoder models are near-isotropic (participation ratio 164 and 95). Standardizing each model's features before RSA removes that dominance and attenuates the reversal from -0.573 to -0.479 (mean-centered) to -0.432 (per-dimension standardized), while encoding is unchanged. We do not read the remainder as a readout-operation residual: the standardized residual's family-clustered CI crosses zero ($[-0.73, +0.23]$), and the new random-direction intervention cannot identify which naturally occurring geometry causes the original reversal. Even standardized, causal models keep the lower RSA and higher encoding (RSA 0.006 versus masked 0.007 versus enc–dec 0.009; encoding 0.049 versus 0.041 versus 0.043), which describes a remaining point-estimate difference, not an identified mechanism: RSA rewards representations whose dominant geometry matches the brain, encoding rewards representations linearly concentrated onto the neural response, and the anisotropy that separates autoregressive from bidirectional models sits on opposite sides of that tradeoff. A principal-component truncation makes the same point descriptively (Figure 10a): encoding peaks at the top few components and then falls, while similarity keeps rising as more of the dominant geometry is added.

Exploratory association within the autoregressive class. Pooled over all 32 checkpoints participation ratio correlates with the encoding$-$RSA gap ($\rho = -0.65$, family-clustered 95% CI $[-0.87, -0.02]$). That interval excludes zero only barely, and the pooled number carries the same between-class inflation as the reversal itself (whose within-class concordance is $\rho = -0.25$, $p = 0.16$), so pooled it is close to a restatement of "is it causal". The test that is not is available inside the one class where participation ratio actually varies. Among the 18 autoregressive checkpoints, over 6 families, the more anisotropic models show the larger readout gap ($\rho = -0.76$, family-clustered 95% CI $[-0.93, -0.09]$ excluding zero; leaving out any single family moves it only within $[-0.85, -0.66]$). Inside the masked and encoder–decoder classes the relation is absent ($+0.07$ and $+0.10$, both n.s.), as expected where participation ratio is near-ceiling. What we cannot claim is a single mechanism spanning all three classes: the pooled correlation survives clustering only marginally and is mostly the between-class offset, so it does not establish a relation that holds *within* each class. Our preregistered prediction, that brain-relevant signal sits in *non-dominant* components, was not supported;

Table 10: Where the readout gap lives, and where dimensionality covaries with it. *Top:* the overall rank correlation is carried by a between-class offset. *Bottom:* per class. Similarity ranks the classes encoder–decoder > masked > causal while encoding ranks them causal > encoder–decoder > masked, which is the whole reversal; the ENC−RSA rank gap separates causal from the rest ($+25.7$ ranks, family-clustered two-sided $p = 0.005$). Participation ratio predicts that gap *inside* the causal class, where it varies, with a family-clustered CI excluding zero; inside the other two classes participation ratio is near-ceiling and the relation is absent. CIs resample model families, not checkpoints; the masked and encoder–decoder classes contain only four and two families, too few for a stable clustered CI, so their point estimates are given without one.

Concordance decomposition (32 checkpoints, 12 families, 3 classes)

overall similarity-vs-encoding concordance	$\rho = -0.573$	$p = 6 \times 10^{-4}$
after removing architecture-class means	$\rho = -0.25$	$p = 0.16$
between-class share of $\lvert\rho\rvert$	56%	

Per architecture class

class	n	fam.	mean RSA	mean ENC	rank gap	PR–gap ρ [family CI]
causal	18	6	0.0049	0.0494	$+11.2$	-0.76 $[-0.93, -0.09]$
encoder–decoder	5	2	0.0101	0.0426	-18.8	$+0.10$ (CI unstable)
masked	9	4	0.0082	0.0412	-12.0	$+0.07$ (CI unstable)

this effective-dimensionality account is exploratory. Table 10 collects the decomposition and the per-class mechanism.

A learned encoder does not resolve it, and why. If the disagreement were an artifact of hand-crafted neural features, a powerful learned encoder should reconcile the two readouts. We trained a diagonal state-space (S4D) encoder on the raw word-locked MEG with a self-supervised masked-reconstruction objective (no language model involved) and used its 128-dimensional embedding as both readouts' neural representation. The concordance collapses to $\rho_z = +0.07$ (n.s.; ρ_z is the headline concordance statistic recomputed with the learned embedding z in place of the raw neural pattern). This is *not* evidence that the reversal was a featurization artifact. A diagnostic shows the learned embedding's word dissimilarity matrix is near-degenerate (pairwise-distance standard deviation 0.005 versus 0.23 for the raw pattern) and uncorrelated with the raw neural RDM ($\rho \approx 0.01$): reconstruction pre-training captured the word-*general* evoked response and discarded word-*discriminative* geometry, so both readouts lose their ability to rank models (RSA ranking reliability falls $0.98 \rightarrow 0.61$). The test is therefore uninformative rather than a refutation, and it carries a caution of its own: reconstruction-trained neural encoders can flatten exactly the fine word-level structure that RSA-based brain–model alignment depends on.

P Leaderboard Reliability in ML

The instabilities we measure have direct ML analogues: benchmark conclusions can depend on the arbitrary composition of the suite (Dehghani et al., 2021), small evaluation sets leave model comparisons underpowered (Card et al., 2020), and evaluation frameworks respond by making coverage and uncertainty explicit (Liang et al., 2023) or by collecting evaluation data dynamically (Kiela et al., 2021). Brain–LLM benchmarks add a variance axis those suites lack: participants and stimuli are *both* samples, and §5.2 shows the conclusion can change depending on which one the error bar resamples.

Documentation, Hosting, Licensing, and Maintenance

What will be released. On publication, the BrainLLM-Bench release will have two parts: (i) the fixed protocol as a runnable harness (band-splitting, regioning, RSA, confound battery, per-band ceilings, clustered nulls) with a datasheet documenting each cohort's provenance, size, and preprocessing; and (ii) the precomputed alignment tensors, the per-cohort headline cells for the four fully controlled cohorts, and the scorecard, so a new model can be ranked without recomputing the existing

Cohort	Modality	Language	Recordings	Subjects/patients
LibriBrain (Özdogan et al., 2025)	MEG	English	36	1
MEG-MASC (Gwilliams et al., 2023)	MEG	English	76	10 (27 rebuilt)
MEG-SCANS (Habersetzer et al., 2025)	MEG	German	92	23 (of 24)
Brennan (Brennan & Hale, 2019)	EEG	English	n/a	49
ECoG ds005574 (Zada et al., 2025)	ECoG	English	n/a	9

Table 11: The five benchmark cohorts. Modality spans MEG, scalp EEG, and intracranial ECoG; language spans English and German. The task is natural language comprehension throughout.

entries. For review, an anonymized subset is provided in the supplementary material: the harness, its stored-result checks, the datasheet, the derived result files and the sensor-region atlases. The missing baseline tensors and raw-data rerun requirements are disclosed above; public hosting and deposition remain as described below.

Access and licensing. The harness will be released under a permissive open-source license; derived tensors and tables will be released under the terms of the recordings they are computed from, so those derived from LibriBrain (CC BY-NC 4.0) are non-commercial. The underlying recordings keep their original licenses and are obtained from source: LibriBrain (CC BY-NC 4.0), MEG-MASC (CC0), MEG-SCANS (OpenNeuro `ds006468`, CC0) and the "Podcast" ECoG dataset (OpenNeuro `ds005574`, CC0) are public, the Brennan EEG cohort is distributed by its authors through a university data repository under CC BY 4.0, per its DataCite record (Brennan, 2018), and the fMRI cohort ds003020 is on OpenNeuro (CC0) (LeBel et al., 2026). We will redistribute derived word lists, region maps and alignment tensors, not raw recordings, following each source's redistribution and attribution terms.

Hosting and maintenance (release plan). On publication, the code and precomputed tensors will be hosted in a public version-controlled repository with a tagged release matching this paper, and that release will additionally be deposited in a persistent archive with a citable DOI (Zenodo), so the artifact remains resolvable independent of the code host. The datasheet, leaderboard, and scorecard will be versioned alongside the code under semantic versioning, so every reported number is pinned to both a commit and a release tag, and the precomputed tensors will ship with standardized dataset metadata (Croissant) for machine-readable discovery. New models and cohorts are intended to enter through the harness of Section 4.3: a submission runs the fixed protocol, emits the same tensor schema, and is added to the leaderboard through a versioned pull-request review that checks protocol conformance before merge, so growth does not silently change the scoring. The leaderboard and scorecard are the intended points of community growth, and the review checklist and issue tracker will be public so that maintenance is not tied to a single person.

Datasheet (Summary)

Following the datasheet structure of Gebru et al. (2021), the release documents each cohort along the standard axes; a condensed form appears in Table 12. *Motivation*: to make brain–LLM alignment results comparable under one protocol. *Composition*: five cohorts, per-word neural response patterns crossed with per-layer model features, with lexical confounds and reliability ceilings attached. *Collection*: inherited from each source dataset's published acquisition. *Preprocessing*: the fixed protocol of Section 4.2 (band-splitting, regioning, RSA, confound partialization, per-band ceilings, clustered nulls). *Uses*: benchmarking new models and cohorts, auditing prior claims. *Distribution and maintenance*: as in the documentation, hosting, licensing, and maintenance section above, under an open-source license (code) and each source's terms (derived tensors; non-commercial for LibriBrain), versioned to this paper's release.

Historical Scorecard: Operationalizations and Legacy Labels

Table 13 expands each row of the ten-claim scorecard (§L, Table 8) into the operationalization the benchmark ran, its results and original labels. These legacy labels are retained for provenance; Table 8 gives the current scoped outcomes. They do not imply direct replication.

Cohort	Modality	Language	Subjects	Sensors
LibriBrain	MEG	English	1 (12 sessions; 36 recordings)	306
MEG-MASC	MEG	English	10[b]	208
MEG-SCANS	MEG	German	23	306
Brennan	EEG	English	49	60[c]
ECoG (ds005574)	intracranial	English	9	grid (MNI ROIs)

Table 12: Datasheet summary of the five cohorts. Four (LibriBrain, MEG-MASC, MEG-SCANS, ECoG) are scored under the full confound-controlled RSA protocol; the Brennan EEG cohort enters the scorecard and the fixed-depth readout comparison. [b]The tensor build uses the 10-subject preprocessed archive; the stimulus-generalization and concordance analyses (§5.2, §5.3) use the full 27-subject cohort rebuilt from raw, and the text names which build each result uses. [c]61-electrode cap; 60 channels enter the stored epochs and every analysis.

Table 13: Historical scorecard detail: tested operationalizations, stored results and legacy labels. Current scoped outcomes and missing-data status are in Table 8; legacy labels are not replication verdicts (§L); source `results/scorecard/scorecard.tsv`, whose operationalization strings ("7-LLM", "5-cohort") predate the counts used here, which follow the receipts (six tensor-route LLMs in `headline_layer_scan.tsv`; four boundary cohorts in `exp_boundary_conditional/`).

Prior claim (venue)	Benchmark operationalization	Result	Legacy label
Tang et al. (Nat. Neurosci.)	Trial-level contrastive retrieval LLM ← MEG on the LibriBrain headline cell (superseding a naive ridge decoder that had read as null, $R^2 = -0.004$).	Refines: a weaker relative transfers — retrieval clears chance on all seven targets (six models, GPT-2 at two segmentations; 5.7–$6.5\times$ at top-10; shuffled-target $0.7\times$) while absolute accuracy stays far below continuous reconstruction (top-1 0.03%, median rank 2143 of 9495). Modality, temporal-resolution and task-family boundary added: theirs is open-ended generation, ours closed-set retrieval, the contrast their own abstract draws.	refine

continued on next page

Table 13 (continued)

Prior claim (venue)	Benchmark operationalization	Result	Legacy label
Toneva & Wehbe (NeurIPS)	Per-(region, LLM, layer) headline-cell scan across 5 cohorts with the six tensor-route LLMs.	Their claim as stated — middle layers beating both shallower and deeper — holds in 2 of the 8 recomputable cells; the weaker reading (mid-to-shallow over deep) holds on all 4 MEG/EEG cohorts: mid-to-shallow LLM layers carry the headline cell on LibriBrain, MEG-MASC, MEG-SCANS, and Brennan, though the peak latency is cohort-specific ($\approx$346 ms on LibriBrain, 467–768 ms on MEG-MASC, 527–587 ms on MEG-SCANS). ECoG refines the picture: deep layers win on the ECoG headline cell, so the mid-layer story is conditional; their own Fig. 4 offers context length as a competing account of that flip (their deepest layers peak below ten words of context, ours is sentence context).	refine
Schrimpf et al. (PNAS)	Cross-cohort headline z(rho) for the tensor-route LLMs vs random + GloVe + phonemic baselines (LibriBrain, MEG-MASC, MEG-SCANS, Brennan, ECoG).	Replicates qualitatively: LLMs $\gg$ random (historical noise floor, not recomputable), contextual > GloVe in cell-count breadth + layer-depth gradient. But GloVe-300 ties GPT-2 on raw headline magnitude (6.7 vs 6.1, $\times 10^{-3}$; historical, unverified), so the contextual-over-static advantage holds in breadth, not single-cell magnitude; the accuracy–brain-score correlation that is their actual claim is not re-measured here.	refine
Caucheteux et al. (Nat. Hum. Behav.)	Lag analogue of their forecast test (they augment present-word activations with predictions spanning multiple timescales and map onto fMRI in 304 listeners; we shift the model's word sequence against the brain's on the LibriBrain headline cell and locate the peak shift, per post-onset window).	Anticipation is windowed, not general: GPT-2 peaks $+1$ word ahead only in the 350–550 ms windows, peaks *behind* at 150–250 ms, and peaks at zero lag when collapsed over time; BERT shows no forward peak at any window.	refine

continued on next page

Table 13 (continued)

Prior claim (venue)	Benchmark operationalization	Result	Legacy label
Goldstein et al. (Nat. Commun.)	Their claim: in ECoG high-gamma, layer index tracks peak-encoding lag, strongest in Broca's area, for GPT-2 XL and Llama-2. Our analogue: per-(region, LLM, band) layer × peak-time Spearman across 5 cohorts.	Reproduced for BERT temporal_R (r=+0.73 broadband, +0.82 theta) — an architecture they did not test — while the GPT family, their own model class, inverts in the delta band (r=-0.85 to -0.91). Their claim was never stated across frequency bands; adding that dimension, the hierarchy holds neither uniformly across bands nor across architectures.	refine
Zou et al. (Nat. Neurosci.)	4-cohort × 6-LLM × 2-band boundary-conditional RSA; incl. reverse direction with German LLMs.	Consistent with Zou et al. (2026) in 7/12 English cells. In German the effect is model-dependent, not uniformly reversed: multilingual BERT reverses significantly (7/23, p=0.005), German-BERT is null (p=0.96) and German-GPT-2 only trends negative (p=0.09), so we record a model-conditional instability rather than a language mechanism.	refine
Bhaya-Grossman et al. (Nature)	Layer-split contrast multilingual-BERT vs German-native LLMs across LLM depths on 6 subjects of the 23-subject MEG-SCANS cohort.	Multilingual-BERT wins at shallow layers, German-BERT at deep layers. Their split — shared acoustic–phonetic encoding versus experience-dependent word-boundary, word-frequency and phonotactic encoding, reported *within the same STG neural populations*, not across regions — appears here as a split along LLM depth. Cross-modality and cross-language analogue only: their cohort is ECoG in Spanish, English and Mandarin.	**replicate**
Hadidi et al. (Nat. Commun.)	Their confounds, positional signals and word rate, as a confound-partialling battery (three blocks, then an extended six: word rate, position, length, log frequency, lag-1 MEG components, chronology) per cohort and per subject; their datasets are fMRI and ECoG (reading passages or sentences, and listening to stories), ours MEG listening.	Pooled BERT on LibriBrain retains 47% under the extended battery (95% under the first); the pooled GPT family drops to 0%, but per-subject MEG-MASC GPT-2 retains 36% in 6/10 subjects. The confound critique applies, but the signal is not removed under per-subject disaggregation. Static word embeddings, the dominant confound for trained models in Feghhi et al. (2024), are not in this battery.	refine

continued on next page

Table 13 (continued)

Prior claim (venue)	Benchmark operationalization	Result	Legacy label
Antonello et al. (NeurIPS)	Their claim: fMRI encoding performance scales logarithmically with size, 125M–30B, for unidirectional models (OPT, LLaMA). Like-for-like arm: the registered Qwen 0.5B–72B encoding ladder, plus the six-point Pythia 70M→2.8B ladder scored per session at the pre-registered depth; the BERT/GPT-2 RSA comparison is a secondary arm outside their model class.	On the registered bounded-alpha grid, with the rungs re-extracted after the tokenization defect, the ordering reverses on LibriBrain (encoding $\rho = -0.96$; 72B vs 0.5B $p = 0.0005$), but a post hoc training-fold-expanded grid attenuates the size correlation to -0.464 while retaining a 72B deficit; on MEG-MASC the encoding arm fails its rankability gate and is not determined; under the similarity readout size orders positively ($\rho = +0.93$, $p = 0.001$), so the verdict is stated for encoding only. Outside their model class, BERT-base (110M, masked) leads at 13.1×10^{-3}, above the three-times-larger GPT-2-medium (355M, 6.9×10^{-3}); across the ladder size does not predict alignment ($\rho = -0.14$, $p = 0.79$) and 2.8B sits nominally below 70M at forty times the parameters ($p = 0.23$). In these ladders parameter count did not order alignment.	**contradict**
He et al. (NeurIPS)	Iterative residual disentanglement (lexicon, syntax, meaning, reasoning; they test one model family) on LibriBrain, then the reasoning residual split medial versus constituent-onset, read out as alignment per model.	The medial-minus-onset contrast of the *reasoning* residual sign-inverts between architectures (LibriBrain: GPT-2 $z = +13.1 \times 10^{-3}$, BERT $z = -15.7 \times 10^{-3}$). The *meaning* component also inverts but in the opposite direction and roughly fifteen times smaller (GPT-2 -0.7, BERT $+0.9 \times 10^{-3}$), so the architecture-conditional effect is carried by the boundary term, not the semantic one.	refine